Université Aix Marseille 1

Licence de mathématiques

Cours d'Analyse numérique

Raphaèle Herbin

20 août 2010

Table des matières

1 Systèmes linéaires **4**
- 1.1 Objectifs . . . 4
- 1.2 Quelques rappels d'algèbre linéaire . . . 4
  - 1.2.1 Norme induite . . . 4
  - 1.2.2 Rayon spectral . . . 6
  - 1.2.3 Matrices diagonalisables . . . 7
- 1.3 Les méthodes directes . . . 9
  - 1.3.1 Définition . . . 9
  - 1.3.2 Méthode de Gauss et méthode LU . . . 9
  - 1.3.3 Version programmable des algorithmes de Gauss et LU pour un système carré . . . 12
  - 1.3.4 Méthode de Choleski . . . 14
  - 1.3.5 Quelques propriétés . . . 19
- 1.4 Conditionnement . . . 21
  - 1.4.1 Le problème des erreurs d'arrondis . . . 21
  - 1.4.2 Conditionnement et majoration de l'erreur d'arrondi . . . 21
  - 1.4.3 Discrétisation d'équations différentielles, conditionnement "efficace" . . . 23
- 1.5 Méthodes itératives . . . 25
  - 1.5.1 Origine des systèmes à résoudre . . . 25
  - 1.5.2 Définition et propriétés . . . 28
  - 1.5.3 Méthodes de Jacobi, Gauss-Seidel et SOR/SSOR . . . 30
  - 1.5.4 Recherche de valeurs propres et vecteurs propres . . . 36
- 1.6 Exercices . . . 36
- 1.7 Suggestions . . . 56
- 1.8 Corrigés . . . 60

2 Systèmes non linéaires **84**
- 2.1 Les méthodes de point fixe . . . 84
  - 2.1.1 Point fixe de contraction . . . 84
  - 2.1.2 Point fixe de monotonie . . . 87
  - 2.1.3 Vitesse de convergence . . . 89
- 2.2 Méthode de Newton . . . 91
  - 2.2.1 Construction et convergence de la méthode . . . 91
  - 2.2.2 Variantes de la méthode de Newton . . . 95
- 2.3 Exercices . . . 98
- 2.4 Suggestions . . . 105
- 2.5 Corrigés . . . 108

3 Optimisation **122**
3.1 Définitions et rappels . . . 122
3.1.1 Définition des problèmes d'optimisation . . . 122
3.1.2 Rappels et notations de calcul différentiel . . . 122
3.2 Optimisation sans contrainte . . . 123
3.2.1 Définition et condition d'optimalité . . . 123
3.2.2 Résultats d'existence et d'unicité . . . 124
3.3 Algorithmes d'optimisation sans contrainte . . . 128
3.3.1 Méthodes de descente . . . 128
3.3.2 Algorithmes du gradient conjugué . . . 131
3.3.3 Méthodes de Newton et Quasi–Newton . . . 138
3.3.4 Résumé sur les méthodes d'optimisation . . . 141
3.4 Optimisation sous contraintes . . . 142
3.4.1 Définitions . . . 142
3.4.2 Existence – Unicité – Conditions d'optimalité simple . . . 142
3.4.3 Conditions d'optimalité dans le cas de contraintes égalité . . . 143
3.4.4 Contraintes inégalités . . . 146
3.5 Algorithmes d'optimisation sous contraintes . . . 147
3.5.1 Méthodes de gradient avec projection . . . 147
3.5.2 Méthodes de dualité . . . 149
3.6 Exercices . . . 153
3.7 Suggestions . . . 165
3.8 Corrigés . . . 168

4 Equations différentielles **186**
4.1 Introduction . . . 186
4.2 Consistance, stabilité et convergence . . . 189
4.3 Théorème général de convergence . . . 190
4.4 Exemples . . . 193
4.5 Explicite ou implicite ? . . . 194
4.5.1 L'implicite gagne... . . . 194
4.5.2 L'implicite perd... . . . 195
4.5.3 Match nul . . . 195
4.6 Etude du schéma d'Euler implicite . . . 196
4.7 Exercices . . . 197
4.8 Corrigés . . . 206

Introduction

L'objet de l'analyse numérique est de concevoir et d'étudier des méthodes de résolution de certains problèmes mathématiques, en général issus de la modélisation de problèmes "réels", et dont on cherche à calculer la solution à l'aide d'un ordinateur.
Le cours est structuré en quatre grands chapitres :
– Systèmes linéaires
– Systèmes non linéaires
– Optimisation
– Equations différentielles.
On pourra consulter les ouvrages suivants pour ces différentes parties (ceci est une liste non exhaustive !) :
– P.G. Ciarlet, Introduction à l'analyse numérique et à l'optimisation, Masson, 1982, (pour les chapitre 1 à 3 de ce polycopié).
– M. Crouzeix, A.L. Mignot, Analyse numérique des équations différentielles, Collection mathématiques appliquées pour la maitrise, Masson, (pour le chapitre 4 de ce polycopié).
– J.P. Demailly, Analyse numérique et équations différentielles Collection Grenoble sciences Presses Universitaires de Grenoble
– L. Dumas, Modélisation à l'oral de l'agrégation, calcul scientifique, Collection CAPES/Agrégation, Ellipses, 1999.
– J. Hubbard, B. West, Equations différentielles et systèmes dynamiques, Cassini.
– P. Lascaux et R. Théodor, Analyse numérique matricielle appliquée à l'art de l'ingénieur, tomes 1 et 2, Masson, 1987
– L. Sainsaulieu, Calcul scientifique cours et exercices corrigés pour le 2ème cycle et les éécoles d'ingénieurs, Enseignement des mathématiques, Masson, 1996.
– M. Schatzman, Analyse numérique, cours et exercices, (chapitres 1,2 et 4).
– D. Serre, Les matrices, Masson, (2000). (chapitres 1,2 et 4).
– P. Lascaux et R. Theodor, Analyse numérique sappliquée aux sciences de l'ingénieur, Paris, (1994)
– R. Temam, Analyse numérique, Collection SUP le mathématicien, Presses Universitaires de France, 1970.
Et pour les anglophiles...
– M. Braun, Differential Equations and their applications, Springer, New York, 1984 (chapitre 4).
– G. Dahlquist and A. Björck, Numerical Methods, Prentice Hall, Series in Automatic Computation, 1974, Englewood Cliffs, NJ.
– R. Fletcher, Practical methods of optimization, J. Wiley, New York, 1980 (chapitre 3).
– G. Golub and C. Van Loan, Matrix computations, The John Hopkins University Press, Baltimore (chapitre 1).
– R.S. Varga, Matrix iterative analysis, Prentice Hall, Englewood Cliffs, NJ 1962.
Ce cours a été rédigé pour la licence de mathématiques par télé-enseignement de l'université d'Aix-Marseille 1. Chaque chapitre est suivi d'un certian nombre d'exercices. On donne ensuite des suggestions pour effectuer les exercices, puis des corrigés détaillés. Il est fortement conseillé d'essayer de faire les exercices d'abord sans ces indications, et de ne regarder les corrigés détaillés qu'une fois l'exercice achevé (même si certaines questions n'ont pas pu être effectuées), ceci pour se préparer aux conditions d'examen.

Chapitre 1

Systèmes linéaires

1.1 Objectifs

On note $\mathcal{M}_N(\mathbb{R})$ l'ensemble des matrices carrées d'ordre N. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^N$, on a comme objectif de résoudre le système linéaire $Ax = b$, c'est à dire de trouver x solution de :

$$\left\{ \begin{array}{l} x \in \mathbb{R}^N \\ Ax = b \end{array} \right. \tag{1.1.1}$$

Comme A est inversible, il existe un unique vecteur $x \in \mathbb{R}^N$ solution de (1.1.1). Nous allons étudier dans les deux chapitres suivants des méthodes de calcul de ce vecteur x : la première partie de ce chapitre sera consacrée aux méthodes "directes" et la deuxième aux méthodes "itératives". Nous aborderons ensuite en troisième partie les méthodes de résolution de problèmes aux valeurs propres.
Un des points essentiels dans l'efficacité des méthodes envisagées concerne la taille des systèmes à résoudre. Entre 1980 et 2000, la taille de la mémoire des ordinateurs a augmenté de façon drastique. La taille des systèmes qu'on peut résoudre sur ordinateur a donc également augmenté, selon l'ordre de grandeur suivant :

1980 :	matrice "pleine" (tous les termes sont non nuls)	$N = 10^2$
	matrice "creuse"	$N = 10^6$
2000 :	matrice "pleine"	$N = 10^6$
	matrice "creuse"	$N = 10^8$

Le développement des méthodes de résolution de systèmes linéaires est liée à l'évolution des machines informatiques. Un grand nombre de recherches sont d'ailleurs en cours pour profiter au mieux de l'architecture des machines (méthodes de décomposition en sous domaines pour profiter des architectures parallèles, par exemple).
Dans la suite de ce chapitre, nous verrons deux types de méthodes pour résoudre les systèmes linéaires : les méthodes directes et les méthodes itératives. Pour faciliter la compréhension de leur étude, nous commençons par quelques rappels d'algèbre linéaire.

1.2 Quelques rappels d'algèbre linéaire

1.2.1 Norme induite

Définition 1.1 (Norme matricielle, norme induite) *On note $\mathcal{M}_N(\mathbb{R})$ l'espace vectoriel (sur $\mathbb{R}$) des matrices carrées d'ordre N.*

1. *On appelle norme matricielle sur $\mathcal{M}_N(\mathbb{R})$ une norme $\|\cdot\|$ sur $\mathcal{M}_N(\mathbb{R})$ t.q.*

$$\|AB\| \leq \|A\|\|B\|, \ \forall A, B \in \mathcal{M}_N(\mathbb{R}) \tag{1.2.2}$$

2. *On considère $\mathbb{R}^N$ muni d'une norme $\|\cdot\|$. On appelle norme matricielle induite (ou norme induite) sur $\mathcal{M}_N(\mathbb{R})$ par la norme $\|\cdot\|$, encore notée $\|\cdot\|$, la norme sur $\mathcal{M}_N(\mathbb{R})$ définie par :*

$$\|A\| = \sup\{\|Ax\|; \quad x \in \mathbb{R}^N, \|x\| = 1\}, \ \forall A \in \mathcal{M}_N(\mathbb{R}) \tag{1.2.3}$$

Proposition 1.2 *Soit $\mathcal{M}_N(\mathbb{R})$ muni d'une norme induite $\|\cdot\|$. Alors pour toute matrice $A \in \mathcal{M}_N(\mathbb{R})$, on a :*

1. $\|Ax\| \leq \|A\| \ \|x\|, \quad \forall x \in \mathbb{R}^N$,
2. $\|A\| = \max\left\{\|Ax\| \ ; \ \|x\| = 1, \ x \in \mathbb{R}^N\right\}$,
3. $\|A\| = \max\left\{\dfrac{\|Ax\|}{\|x\|} \ ; \ x \in \mathbb{R}^N \setminus \{0\}\right\}$.
4. $\|\cdot\|$ *est une norme matricielle*.

Démonstration :

1. Soit $x \in \mathbb{R}^N \setminus \{0\}$, posons $y = \dfrac{x}{\|x\|}$, alors $\|y\| = 1$ donc $\|Ay\| \leq \|A\|$. On en déduit $\dfrac{\|Ax\|}{\|x\|} \leq \|A\|$ et donc $\|Ax\| \leq \|A\| \ \|x\|$. Si maintenant $x = 0$, alors $Ax = 0$, et donc $\|x\| = 0$ et $\|Ax\| = 0$; l'inégalité $\|Ax\| \leq \|A\| \ \|x\|$ est encore vérifiée.
2. L'application φ définie de $\mathbb{R}^N$ dans $\mathbb{R}$ par : $\varphi(x) = \|Ax\|$ est continue sur la sphère unité $S_1 = \{x \in \mathbb{R}^N \mid \|x\| = 1\}$ qui est un compact de $\mathbb{R}^N$. Donc φ est bornée et atteint ses bornes : il existe $x_0 \in \mathbb{R}^N$ tel que $\|A\| = \|Ax_0\|$
3. La dernière égalité résulte du fait que $\dfrac{\|Ax\|}{\|x\|} = \|A\dfrac{x}{\|x\|}\|$ et $\dfrac{x}{\|x\|} \in S_1$ pour tout $x \neq 0$.

■

Proposition 1.3 *Soit $A = (a_{i,j})_{i,j\in\{1,\ldots,N\}} \in \mathcal{M}_N(\mathbb{R})$.*

1. *On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Alors*

$$\|A\|_\infty = \max_{i\in\{1,\ldots,N\}} \sum_{j=1}^{N} |a_{i,j}|. \tag{1.2.4}$$

2. *On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Alors*

$$\|A\|_1 = \max_{j\in\{1,\ldots,N\}} \sum_{i=1}^{N} |a_{i,j}| \tag{1.2.5}$$

3. *On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_2$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_2$.*

$$\|A\|_2 = (\rho(A^tA))^{\frac{1}{2}}. \tag{1.2.6}$$

La démonstration de cette proposition fait l'objet de l'exercice 3 page 36

1.2.2 Rayon spectral

Définition 1.4 (Valeurs propres et rayon spectral) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On appelle valeur propre de A tout $\lambda \in \mathbb{C}$ tel qu'il existe $x \in \mathbb{C}^N$, $x \neq 0$ tel que $Ax = \lambda x$. L'élément x est appelé vecteur propre de A associé à λ. On appelle rayon spectral de A la quantité $\rho(A) = \max\{|\lambda|;\ \lambda \in \mathbb{C},\ \lambda$ valeur propre de $A\}$.*

Lemme 1.5 (Convergence et rayon spectral) *On munit $\mathcal{M}_N(\mathbb{R})$ d'une norme, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_N(\mathbb{R})$. Alors :*

1. *$\rho(A) < 1$ si et seulement si $A^k \to 0$ quand $k \to \infty$.*
2. *$\rho(A) < 1 \Rightarrow \limsup_{k\to\infty} \|A^k\|^{\frac{1}{k}} \leq 1$.*
3. *$\liminf_{k\to\infty} \|A^k\|^{\frac{1}{k}} < 1 \Rightarrow \rho(A) < 1$.*
4. *$\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}$.*
5. *On suppose de plus que $\|\cdot\|$ une norme matricielle (induite ou non). Alors*

$$\rho(A) \leq \|A\|.$$

La démonstration de ce lemme fait l'objet de l'exercice 5 page 37. Elle nécessite un résultat d'approximation du rayon spectral par une norme induite bien choisie, que voici :

Remarque 1.6 (Convergence des suites) *Une conséquence immédiate du lemme 1.5 page 6 est que si $x^{(0)}$ est donné et $x^{(k)}$ défini par $x^{(k+1)} = Ax^{(k)}$, alors la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers 0 si et seulement si $\rho(A) < 1$.*

Proposition 1.7 (Rayon spectral et norme induite) *Soient $A \in \mathcal{M}_N(\mathbb{R})$ et $\varepsilon > 0$. Il existe une norme sur $\mathbb{R}^N$ (qui dépend de A et ε) telle que la norme induite sur $\mathcal{M}_N(\mathbb{R})$, notée $\|\cdot\|_{A,\varepsilon}$, vérifie $\|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon$.*

Démonstration :
Soit $A \in \mathcal{M}_N(\mathbb{R})$, alors par le lemme 1.8 donné ci-après, il existe une base $(f_1, \ldots, f_N)$ de $\mathbb{C}^N$ et une famille de complexes $(\lambda_{i,j})_{i,j=1,\ldots,N, j\leq i}$ telles que $Af_i = \sum_{j\leq i} \lambda_{i,j} f_j$. Soit $\eta \in]0,1[$, pour $i = 1, \ldots, N$, on définit $e_i = \eta^{i-1} f_i$. La famille $(e_i)_{i=1,\ldots,N}$ forme une base de $\mathbb{C}^N$. On définit alors une norme sur $\mathbb{R}^N$ par $\|x\| = (\sum_{i=1}^N \alpha_i \overline{\alpha_i})^{1/2}$, où les α_i sont les composantes de x dans la base $(e_i)_{i=1,\ldots,N}$. Notons que cette norme dépend de A et de η. Soit $\varepsilon > 0$. Montrons que si η bien choisi, on a $\|A\| \leq \rho(A) + \varepsilon$. Soit $x = \sum_{i=1,\ldots,N} \alpha_i e_i$. Comme

$$Ae_i = \sum_{1\leq j\leq i} \eta^{i-j} \lambda_{i,j} e_j,$$

on a donc :

$$Ax = \sum_{i=1}^N \sum_{1\leq j\leq i} \eta^{i-j} \lambda_{i,j} \alpha_i e_j = \sum_{j=1}^N \Big(\sum_{i=j}^N \eta^{i-j} \lambda_{i,j} \alpha_i\Big) e_j$$

On en déduit que

$$\|Ax\|^2 = \sum_{j=1}^N \Big(\sum_{i=j}^N \eta^{i-j} \lambda_{i,j} \alpha_i\Big) \Big(\sum_{i=j}^N \eta^{i-j} \overline{\lambda}_{i,j} \overline{\alpha}_i\Big),$$

soit encore

$$\|Ax\|^2 = \sum_{j=1}^N \lambda_{j,j} \overline{\lambda_{j,j}} \alpha_j \overline{\alpha_j} + \sum_{j=1}^N \sum_{\substack{k,\ell\geq j \\ (k,\ell)\neq(j,j)}} \eta^{k+\ell-2j} \lambda_{k,j} \overline{\lambda_{\ell,j}} \alpha_k \overline{\alpha_\ell}.$$

Comme $\eta \in]0,1[$, on en conclut que :

$$\|Ax\|^2 \leq \rho(A)\|x\|^2 + N^3 \rho(A)^2 \|x\|^2 \eta.$$

D'où le résultat, en prenant η tel que $\eta N^3 \rho(A)^2 < \varepsilon$.

■

Lemme 1.8 (Triangularisation d'une matrice) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée quelconque, alors il existe une base $(f_1, \ldots, f_N)$ de $\mathbb{C}$ et une famille de complexes $(\lambda_{i,j})_{i=1,\ldots,N, j=1,\ldots,N, j<i}$ telles que $Af_i = \lambda_{i,i} f_i + \sum_{j<i} \lambda_{i,j} f_j$. De plus $\lambda_{i,i}$ est valeur propre de A pour tout $i \in \{1, \ldots, N\}$.*

On admettra ce lemme.
Nous donnons maintenant un théorème qui nous sera utile dans l'étude du conditionnement, ainsi que plus tard dans l'étude des méthodes itératives.

Théorème 1.9 (Matrices de la forme $Id + A$)

1. *Soit une norme matricielle induite, Id la matrice identité de $\mathcal{M}_N(\mathbb{R})$ et $A \in \mathcal{M}_N(\mathbb{R})$ telle que $\|A\| < 1$. Alors la matrice $Id + A$ est inversible et*

$$\|(Id + A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

2. *Si une matrice de la forme $Id + A \in \mathcal{M}_N(\mathbb{R})$ est singulière, alors $\|A\| \geq 1$ pour toute norme matricielle $\|\cdot\|$.*

Démonstration :

1. La démonstration du point 1 fait l'objet de l'exercice 9 page 37.
2. Si la matrice $Id + A \in \mathcal{M}_N(\mathbb{R})$ est singulière, alors $\lambda = -1$ est valeur propre, et donc en utilisant la proposition 1.7, on obtient que $\|A\| \geq \rho(A) \geq 1$.

■

1.2.3 Matrices diagonalisables

Définition 1.10 (Matrice diagonalisable) *Soit A une matrice réelle carrée d'ordre n. On dit que A est diagonalisable dans $\mathbb{R}$ si il existe une base $(\phi_1, \ldots, \phi_n)$ et des réels $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\phi_i = \lambda_i \phi_i$ pour $i = 1, \ldots, n$. Les réels $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de A, et les vecteurs $\phi_1, \ldots, \phi_n$ sont les vecteurs propres associés.*

Vous connaissez sûrement aussi la diagonalisation dans $\mathbb{C}$: une matrice réelle carrée d'ordre n est diagonalisable dans $\mathbb{C}$ si il existe une base $(\phi_1, \ldots, \phi_n)$ de $\mathbb{C}^n$ et des nombres complexes $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\phi_i = \lambda_i \phi_i$ pour $i = 1, \ldots, n$. Ici et dans toute la suite, comme on résout des systèmes linéaire réels, on préfère travailler avec la diagonlaisation dans $\mathbb{R}$; cependant il y a des cas où la diagonalisation dans $\mathbb{C}$ est utile et même nécessaire (étude de stabilité des systèmes diférentiels, par exemple). Par souci de clarté, nous préciserons toujours si la diagonalisation considérée est dans $\mathbb{R}$ ou dans $\mathbb{C}$.

Lemme 1.11 *Soit A une matrice réelle carrée d'ordre n, diagonalisable dans $\mathbb{R}$. Alors*

$$A = P \mathrm{d}iag(\lambda_1, \ldots, \lambda_n) P^{-1},$$

où P est la matrice dont les vecteurs colonnes sont égaux aux vecteurs $\phi_1, \ldots, \phi_n$.

Démonstration Soit P la matrice définie (de manière unique) par $Pe_i = \phi_i$, où $(e_i)_{i=1,\ldots,n}$ est la base canonique de $\mathbb{R}^n$, c'est-à-dire que $(e_i)_j = \delta_{i,j}$. La matrice P est appelée matrice de passage de la base $(e_i)_{i=1,\ldots,n}$ à la base $(\phi_i)_{i=1,\ldots,n}$; la i-ème colonne de P est constituée des composantes de ϕ_i dans la base canonique $(e_1, \ldots, e_N)$.
La matrice P est évidemment inversible, et on peut écrire :

$$A\phi_i = APe_i = \lambda_i \phi_i,$$

de sorte que :

$$P^{-1}APe_i = \lambda_i P^{-1}\phi_i = \lambda_i e_i.$$

On a donc bien $P^{-1}AP = diag(\lambda_1, \ldots, \lambda_n)$ (ou encore $A = Pdiag(\lambda_1, \ldots, \lambda_n)P^{-1}$).

La diagonalisation des matrices réelles symétriques est un outil qu'on utilisera souvent dans la suite, en particulier dans les exercices.

Lemme 1.12 *Soit E un espace vectoriel sur $\mathbb{R}$ de dimension finie : $dimE = n$, $n \in \mathbb{N}^*$, muni d'un produit scalaire i.e. d'une application*

$$E \times E \to \mathbb{R},$$
$$(x, y) \to (x \mid y)_E,$$

qui vérifie :

$$\forall x \in E, (x \mid x)_E \geq 0 \text{ et } (x \mid x)_E = 0 \Leftrightarrow x = 0,$$
$$\forall (x, y) \in E^2, (x \mid y)_E = (y \mid x)_E,$$
$$\forall y \in E, \text{ l'application de } E \text{ dans } \mathbb{R}, \text{ définie par } x \to (x \mid y)_E \text{ est linéaire.}$$

Ce produit scalaire induit une norme sur E, $\|x\| = \sqrt{(x \mid x)_E}$.
Soit T une application linéaire de E dans E. On suppose que T est symétrique, c.à.d. que $(T(x) \mid y)_E = (x \mid T(y))_E, \forall (x, y) \in E^2$. Alors il existe une base orthonormée $(f_1 \ldots f_n)$ de E (c.à.d. telle que $(f_i, f_j)_E = \delta_{i,j}$) et $(\lambda_1 \ldots \lambda_n) \in \mathbb{R}^n$ tels que $T(f_i) = \lambda_i f_i$ pour tout $i \in \{1 \ldots n\}$.

Conséquence immédiate : Dans le cas où $E = \mathbb{R}^N$, le produit scalaire canonique de $x = (x_1, \ldots, x_N)^t$ et $y = (y_1, \ldots, y_N)^t$ est défini par $(x \mid y)_E = x \cdot y = \sum_{i=1}^N x_i y_i$. Si $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice symétrique, alors l'application T définie de E dans E par : $T(x) = Ax$ est linéaire, et : $(Tx|y) = Ax \cdot y = x \cdot A^t y = x \cdot Ay = (x \mid Ty)$. Donc T est linéaire symétrique. Par le lemme précédent, il existe $(f_1 \ldots f_N)$ et $(\lambda_1 \ldots \lambda_N) \in \mathbb{R}$ tels que $Tf_i = Af_i = \lambda_i f_i \, \forall \, i \in \{1, \ldots, N\}$ et $f_i \cdot f_j = \delta_{i,j}, \forall \, (i, j) \in \{1, \ldots, N\}^2$.

Interprétation algébrique : Il existe une matrice de passage P de $(e_1, \ldots, e_N)$ base canonique dans $(f_1, \ldots, f_N)$ dont la première colonne de P est constituée des coordonnées de f_i dans $(e_1 \ldots e_N)$. On a : $Pe_i = f_i$. On a alors $P^{-1}APe_i = P^{-1}Af_i = P^{-1}(\lambda_i f_i) = \lambda_i e_i = diag(\lambda_1, \ldots, \lambda_N)e_i$, où $diag(\lambda_1, \ldots, \lambda_N)$ désigne la matrice diagonale de coefficients diagonaux $\lambda_1, \ldots, \lambda_N$. On a donc :

$$P^{-1}AP = \begin{bmatrix} \lambda_i & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix} = D.$$

De plus P est orthogonale, *i.e.* $P^{-1} = P^t$. En effet,

$$P^t Pe_i \cdot e_j = Pe_i \cdot Pe_j = (f_i | f_j) = \delta_{i,j} \quad \forall i, j \in \{1 \ldots N\},$$

et donc $(P^t Pe_i - e_i) \cdot e_j = 0 \;\; \forall j \in \{1 \ldots N\} \;\; \forall i \in \{1, \ldots N\}$. On en déduit $P^t Pe_i = e_i$ pour tout $i = 1, \ldots N$, *i.e.* $P^t P = PP^t = Id$.

Démonstration du lemme 1.12 Cette démonstration se fait par récurrence sur la dimension de E.
1ère étape.
On suppose $dimE = 1$. Soit $e \in E, e \neq 0$, alors $E = \mathbb{R}e = f_1$ avec $f_1 = \dfrac{e}{\|e\|}$. Soit $T : E \to E$ linéaire symétrique, on a : $Tf_1 \in \mathbb{R}f_1$ donc il existe $\lambda_1 \in \mathbb{R}$ tel que $Tf_1 = \lambda_1 f_1$.

2ème étape.

On suppose le lemme vrai si $dimE < n$. On montre alors le lemme si $dimE = n$. Soit E un espace vectoriel normé sur $\mathbb{R}$ tel que $dimE = n$ et $T : E \to E$ linéaire symétrique. Soit φ l'application définie par :

$$\begin{array}{rl} \varphi : & E \to \mathbb{R} \\ & x \to (Tx|x). \end{array}$$

L'application φ est continue sur la sphère unité $S_1 = \{x \in E|\ \|x\| = 1\}$ qui est compacte car $dimE < +\infty$; il existe donc $e \in S_1$ tel que $\varphi(x) \leq \varphi(e) = (Te \mid e) = \lambda$ pour tout $x \in E$. Soit $y \in E \setminus \{0\}$, et soit $t \in]0, \frac{1}{\|y\|}[$ alors $e + ty \neq 0$. On en déduit que :

$$\frac{e+ty}{\|e+ty\|} \in S_1 \text{ donc } \varphi(e) = \lambda \geq \left(T\frac{(e+ty)}{\|e+ty\|} | \frac{e+ty}{\|e+ty\|} \right)_E$$

donc $\lambda(e + ty \mid e + ty)_E \geq (T(e + ty) \mid e + ty)$. En développant on obtient :

$$\lambda[2t(e \mid y) + t^2(y \mid y)_E] \geq 2t(T(e) \mid y) + t^2(T(y) \mid y)_E.$$

Comme $t > 0$, ceci donne :

$$\lambda[2(e \mid y) + t(y \mid y)_E] \geq 2(T(e) \mid y) + t(T(y) \mid y)_E.$$

En faisant tendre t vers 0^+, on obtient $2\lambda(e \mid y)_E \geq 2(T(e) \mid y)$, Soit $0 \geq (T(e) - \lambda e \mid y)$ pour tout $y \in E \setminus \{0\}$. De même pour $z = -y$ on a $0 \geq (T(e) - \lambda e|z)$ donc $(T(e) - \lambda e \mid y) \geq 0$. D'où $(T(e) - \lambda e \mid y) = 0$ pour tout $y \in E$. On en déduit que $T(e) = \lambda e$. On pose $f_n = e$ et $\lambda_n = \lambda$.

Soit $F = \{x \in E; (x \mid e) = 0\}$, on a donc $F \neq E$, et $E = F \bigoplus \mathbb{R}e$: on peut décomposer $x \in E$ comme $(x = x - (x \mid e)e + (x \mid e)e)$. L'application $S = T|_F$ est linéaire symétrique et on a $dimF = n - 1$. et $S(F) \subset F$. On peut donc utiliser l'hypothèse de récurrence : $\exists(\lambda_1 \dots \lambda_{n-1}) \in \mathbb{R}^n$ et $\exists(f_1 \dots f_{n-1}) \in E^n$ tels que $\forall\, i \in \{1 \dots n-1\}$, $Sf_i = Tf_i = \lambda_i f_i$, et $\forall i, j \in \{1 \dots n-1\}, (f_i \mid f_j) = \delta_{i,j}$. Et donc $(\lambda_1 \dots \lambda_N)$ et $(f_1 \dots f_N)$ conviennent.

1.3 Les méthodes directes

1.3.1 Définition

Définition 1.13 (Méthode directe) *On appelle méthode directe de résolution de (1.1.1) une méthode qui donne exactement x (A et b étant connus) solution de (1.1.1) après un nombre fini d'opérations élémentaires* $(+, -, \times, /)$.

Parmi les méthodes de résolution du système (1.1.1) on citera :

- la méthode de Gauss (avec pivot)
- la méthode LU, qui est une réécriture de la méthode Gauss.

Nous rappelons la méthode de Gauss et la méthode LU et nous étudierons plus en détails la méthode de Choleski, qui est adaptée aux matrices symétriques.

1.3.2 Méthode de Gauss et méthode LU

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^N$. On cherche à calculer $x \in \mathbb{R}^N$ tel que $Ax = b$. Le principe de la méthode de Gauss est de se ramener, par des opérations simples (combinaisons linéaires), à un système triangulaire équivalent, qui sera donc facile à inverser. On pose $A^{(1)} = A$ et $b^{(1)} = b$. Pour $i = 1, \dots, N-1$, on cherche à calculer $A^{(i+1)}$ et $b^{(i+1)}$ tels que les systèmes $A^{(i)}x = b^{(i)}$ et $A^{(i+1)}x = b^{(i+1)}$ soient équivalents, où $A^{(i)}$ est une matrice de la forme suivante :
Une fois la matrice $A^{(N)}$ (triangulaire supérieure) et le vecteur $b^{(N)}$ calculés, il sera facile de résoudre le système $A^{(N)}x = b^{(N)}$. Le calcul de $A^{(N)}$ est l'étape de "factorisation", le calcul de $b^{(N)}$ l'étape de "descente", et le calcul de x l'étape de "remontée". Donnons les détails de ces trois étapes.

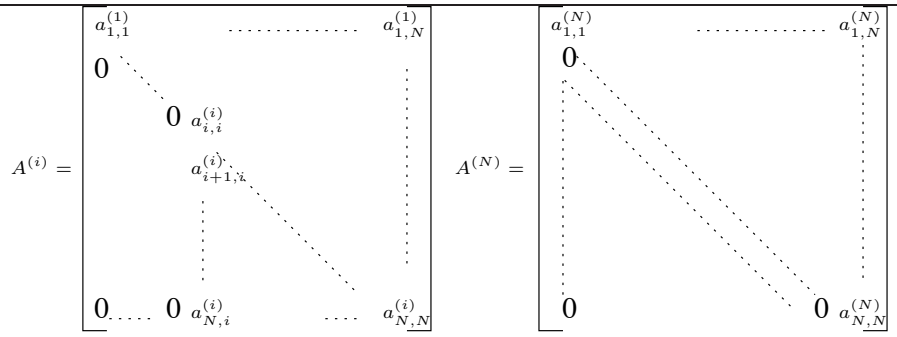

FIGURE 1.1 – Allure des matrices de Gauss à l'étape i et à l'étape N

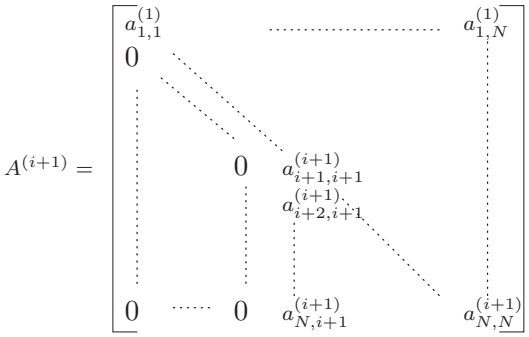

FIGURE 1.2 – Allure de la matrice de Gauss à l'étape $i+1$

Etape de factorisation et descente Pour passer de la matrice $A^{(i)}$ à la matrice $A^{(i+1)}$, on va effectuer des combinaisons linéaires entre lignes qui permettront d'annuler les coefficients de la i-ème colonne situés en dessous de la ligne i (dans le but de se rapprocher d'une matrice triangulaire supérieure). Evidemment, lorsqu'on fait ceci, il faut également modifier le second membre b en conséquence. L'étape de factorisation et descente s'écrit donc :

1. Pour $k \leq i$ et pour $j = 1, \ldots, N$, on pose $a_{k,j}^{(i+1)} = a_{k,j}^{(i)}$ et $b_k^{(i+1)} = b_k^{(i)}$.
2. Pour $k > i$, **si** $a_{i,i}^{(i)} \neq 0$, on pose :

$$a_{k,j}^{(i+1)} = a_{k,j}^{(i)} - \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}} a_{i,j}^{(i)}, \text{ pour } k = j, \ldots, N, \tag{1.3.7}$$

$$b_k^{(i+1)} = b_k^{(i)} - \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}} b_i^{(i)}. \tag{1.3.8}$$

La matrice $A^{(i+1)}$ est de la forme donnée sur la figure 1.3.2. Remarquons que le système $A^{(i+1)}x = b^{(i+1)}$ est bien équivalent au système $A^{(i)}x = b^{(i)}$.
Si la condition $a_{i,i}^{(i)} \neq 0$ est vérifiée pour $i = 1$ à N, on obtient par le procédé de calcul ci-dessus un système linéaire $A^{(N)}x = b^{(N)}$ équivalent au système $Ax = b$, avec une matrice $A^{(N)}$ triangulaire supérieure facile à inverser. On verra un peu plus loin les techniques de pivot qui permettent de régler le cas où la condition $a_{i,i}^{(i)} \neq 0$ n'est pas vérifiée.

Etape de remontée Il reste à résoudre le système $A^{(N)}x = b^{(N)}$. Ceci est une étape facile. Comme $A^{(N)}$ est une matrice inversible, on a $a_{i,i}^{(i)} \neq 0$ pour tout $i1, \ldots, N$, et comme $A^{(N)}$ est une matrice triangulaire supérieure, on peut donc calculer les composantes de x en "remontant", c'est–à–dire de la composante x_N à la composante x_1 :

$$x_N = \frac{b^{(N)}}{a_{N,N}^{(N)}},$$

$$x_i = \frac{1}{a_{i,i}^{(N)}}(b^{(i)} - \sum_{j=i+1,N} a_{i,j}^{(N)} x_j), i = N-1, \ldots, 1.$$

Coût de la méthode de Gauss (nombre d'opérations) On peut montrer (on fera le calcul de manière détaillée pour la méthode de Choleski dans la section suivante, le calcul pour Gauss est similaire) que le nombre d'opérations nécessaires pour effectuer les étapes de factorisation, descente et remontée est $\frac{2}{3}N^3 + O(N^2)$.
En ce qui concerne la place mémoire, on peut très bien stocker les itérés $A^{(i)}$ dans la matrice A de départ, ce qu'on n'a pas voulu faire dans le calcul précédent, par souci de clarté.

Décomposition LU Si le système $Ax = b$ doit être résolu pour plusieurs second membres b, il est évident qu'on a intérêt à ne faire l'étape de factorisation (*i.e.* le calcul de $A^{(N)}$), qu'une seule fois, alors que les étapes de descente et remontée (*i.e.* le calcul de $b^{(N)}$ et x) seront faits pour chaque vecteur b. L'étape de factorisation peut se faire en décomposant la matrice A sous la forme LU.
On admettra le théorème suivant (voir par exemple le livre de Ciarlet cité en début de cours.

Théorème 1.14 (Décomposition LU d'une matrice) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, il existe une matrice de permutation P telle que, pour cette matrice de permutation, il existe un et un seul couple de matrices (L, U) où L est triangulaire inférieure de termes diagonaux égaux à 1 et U est triangulaire supérieure, vérifiant*

$$PA = LU.$$

Cette décomposition peut se calculer très facilement à partir de la méthode de Gauss. Pour simplifier l'écriture, on supposera ici que lors de la méthode de Gauss, la condition $a_{i,i}^{(i)} \neq 0$ est vérifiée pour tout $i = 1, \ldots, N$. Dans ce cas, la matrice de permutation P du théorème 1.14 est la matrice identité. La matrice L a comme coefficients $\ell_{i,j} = \frac{a_{i,j}^{(i)}}{a_{i,i}^{(i)}}$ pour $i > j$, $\ell_{i,i} = 1$ pour tout $i = 1, \ldots, N$, et $\ell_{i,j} = 0$ pour $j > i$, et la matrice U est égale à la matrice $A^{(N)}$. On peut vérifier que $A = LU$ grâce au fait que le système $A^{(N)}x = b^{(N)}$ est équivalent au système $Ax = b$. En effet, comme $A^{(N)}x = b^{(N)}$ et $b^{(N)} = L^{-1}b$, on en déduit que $LUx = b$, et comme A et LU sont inversibles, on en déduit que $A^{-1}b = (LU)^{-1}b$ pour tout $b \in \mathbb{R}^N$. Ceci démontre que $A = LU$.

Techniques de pivot Dans la présentation de la méthode de Gauss et de la décomposition LU, on a supposé que la condition $a_{i,i}^{(i)} \neq 0$ était vérifiée à chaque étape. Or il peut s'avérer que ce ne soit pas le cas, ou que, même si la condition est vérifiée, le "pivot" $a_{i,i}^{(i)}$ soit très petit, ce qui peut entraîner des erreurs d'arrondi importantes dans les calculs. On peut résoudre ce problème en utilisant les techniques de "pivot partiel" ou "pivot total", qui reviennent à choisir une matrice de permutation P qui n'est pas forcément égale à la matrice identité dans le théorème 1.14.
Plaçons-nous à l'itération i de la méthode de Gauss. Comme la matrice $A^{(i)}$ est forcément non singulière, on a :

$$det(A^{(i)}) = a_{1,1}^{(i)} a_{2,2}^{(i)} \cdots a_{i-1,i-1}^{(i)} det \begin{pmatrix} a_{i,i}^{(i)} & \ldots & a_{i,N}^{(i)} \\ \vdots & \ddots & \vdots \\ a_{N,i}^{(i)} & \ldots & a_{N,N}^{(i)} \end{pmatrix} \neq 0.$$

On a donc en particulier

$$det \begin{pmatrix} a_{i,i}^{(i)} & \dots & a_{i,N}^{(i)} \\ \vdots & \ddots & \vdots \\ a_{N,i}^{(i)} & \dots & a_{N,N}^{(i)} \end{pmatrix} \neq 0.$$

Pivot partiel On déduit qu'il existe $i_0 \in \{i, \dots, N\}$ tel que $a_{i_0,i}^{(i)} \neq 0$. On choisit alors $i_0 \in \{i, \dots, N\}$ tel que $|a_{i_0,i}^{(i)}| = \max\{|a_{k,i}^{(i)}|, k = i, \dots, N\}$. On échange alors les lignes i et i_0 (dans la matrice A et le second membre b) et on continue la procédure de Gauss décrite plus haut.

Pivot total On choisit maintenant i_0 et $j_0 \in \{i, \dots, N\}$ tels que $|a_{i_0,j_0}^{(i)}| = \max\{|a_{k,j}^{(i)}|, k = i, \dots, N, j = i, \dots, N\}$, et on échange alors les lignes i et i_0 (dans la matrice A et le second membre b), les colonnes j et j_0 de A et les inconnues x_j et x_{j_0}.

L'intérêt de ces stratégies de pivot est qu'on aboutit toujours à la résolution du système (dès que A est inversible). La stratégie du pivot total permet une moins grande sensibilité aux erreurs d'arrondi. L'inconvénient majeur est qu'on change la structure de A : si, par exemple la matrice avait tous ses termes non nuls sur quelques diagonales seulement, ceci n'est plus vrai pour la matrice $A^{(N)}$.

1.3.3 Version programmable des algorithmes de Gauss et LU pour un système carré

On donne dans ce paragraphe la version "programmable" de l'algorithme de Gauss et de la maéthode LU pour une matrice carrée d'ordre n.
On souhaite résoudre $Ax = b$, avec $A \in M_n(\mathbb{R})$ inversible et un ou plusieurs second membres $b \in \mathbb{R}^n$.

Méthode de Gauss sans pivot

1. (Factorisation et descente) Pour i allant de 1 à n, on effectue les calculs suivants :
   (a) On ne change pas la i-ème ligne
   $u_{i,j} = a_{i,j}$ pour $j = i, \dots, n$,
   $y_i = b_i$
   (b) On change les lignes $i+1$ à n (et le second membre) en utilisant la ligne i. Pour j allant de $i+1$ à n :
   $l_{j,i} = \frac{a_{j,i}}{a_{i,i}}$ (si $a_{i,i} = 0$, prendre la méthode avec pivot partiel)
   pour k allant de $i+1$ à n, $a_{j,k} = a_{j,k} - l_{j,i}a_{i,k}$ (noter que $a_{j,i} = 0$)
   $b_j = b_j - l_{j,i}y_i$
2. (Remontée) On calcule x
   Pour i allant de n à 1,
   $x_i = \frac{1}{u_{i,i}}(y_i - \sum_{j=i+1}^{n} u_{i,j}x_j)$

Méthode LU sans pivot

La méthode LU se déduit de la méthode de Gauss en remarquant simplement que, ayant conservé la matrice L, on peut effectuer les calculs sur b après les calculs sur A. Ce qui donne :

1. (Factorisation) Pour i allant de 1 à n, on effectue les calculs suivants :
   (a) On ne change pas la i-ème ligne
   $u_{i,j} = a_{i,j}$ pour $j = i, \dots, n$,

(b) On change les lignes $i+1$ à n (et le second membre) en utilisant la ligne i. Pour j allant de $i+1$ à n :
$l_{j,i} = \frac{a_{j,i}}{a_{i,i}}$ (si $a_{i,i} = 0$, prendre la méthode avec pivot partiel)
pour k de $i+1$ à n, $a_{j,k} = a_{j,k} - l_{j,i}a_{i,k}$ (noter que $a_{j,i} = 0$)

2. (Descente) On calcule y (avec $Ly = b$)
Pour i allant de 1 à n,
$y_i = b_i - \sum_{k=1}^{i-1} l_{i,k}y_k$ (on a implicitement $l_{i,i} = 1$)
3. (Remontée) On calcule x (avec $Ux = y$)
Pour i allant de n à 1,
$x_i = \frac{1}{u_{i,i}}(y_i - \sum_{j=i+1}^{n} u_{i,j}x_j)$

Méthode LU avec pivot partiel

La méthode LU avec pivot partiel consiste simplement à remarquer que l'ordre dans lequel les équations sont prises n'a aucune importance pour l'algorithme. Au départ de l'algorithme, on se donne la bijection t de $\{1,\dots,n\}$ dans $\{1,\dots,n\}$ définie par $t(i) = i$, et qui va être modfiée au cours de l'algorithme pour tenir compte du choix du pivot. On écrit l'algorithme précédent avec les nouveaux indices de ligne $t(i)$, ce qui donne :

1. (Factorisation) Pour i allant de 1 à n, on effectue les calculs suivants :

(a) Choix du pivot (et de $t(i)$) : on cherche $k \in \{i,\dots,n\}$ t.q. $|a_{t(k),i}| = \max\{|a_{t(l),i}|, l \in \{i,\dots,n\}\}$. On change alors t en prenant $t(i) = t(k)$ et $t(k) = t(i)$.
On ne change pas la ligne $t(i)$
$u_{t(i),j} = a_{t(i),j}$ pour $j = i,\dots,n$,

(b) On modifie les lignes $t(j)$ autres que les lignes $t(1),\dots,t(n)$ (et le second membre), en utilisant la ligne $t(i)$. Donc pour $t(j) \in \{t(i+1),\dots,t(n)\}$ (noter qu'on a uniquement besoin de connaï¿½tre l'ensemble , et pas l'ordre) :
$l_{t(j),i} = \frac{a_{t(j),i}}{a_{t(i),i}}$
pour k de $i+1$ à n, $a_{t(j),k} = a_{t(j),k} - l_{t(j),i}a_{t(i),k}$ (noter que $a_{t(j),i} = 0$)

2. (Descente) On calcule y
Pour i allant de 1 à n,
$y_i = b_{t(i)} - \sum_{j=1}^{i-1} l_{t(j),k}y_k$
3. (Remontée) On calcule x
Pour i allant de n à 1,
$x_i = \frac{1}{u_{t(i),i}}(y_i - \sum_{j=i+1}^{n} u_{t(i),j}x_j)$

NB : On a change l'ordre dans lequel les équations sont considérées (le tableau t donne cet ordre). Donc, on a aussi changé l'ordre dans lequel interviennent les composantes du second membre. Par contre, on n'a pas touché à l'ordre dans lequel interviennent les composantes de x et y.

Il reste maintenant à signaler le "miracle" de cet algorithme...Il est inutile de connaitre complètement t pour faire cet algorithme. A l'étape i de l'item 1 (et d'ailleurs aussi à l'étape i de l'item 2), il suffit de connaître $t(j)$ pour j allant de 1 à i, les opérations de $1(b)$ se faisant alors sur toutes les autres lignes (dans un ordre quelconque). Il suffit donc de partir d'une bijection arbitraire de $\{1,\dots,n\}$ dans $\{1,\dots,n\}$ (par exemple l'identité) et de la modifier à chaque étape. Pour que l'algorithme aboutisse, il suffit que $a_{t(i),i} \neq 0$ (ce qui toujours possible car A est inversible). Pour minimiser des erreurs d'arrondis, on a intérêt à choisir $t(i)$ pour que $|a_{t(i),i}|$ soit le plus grand possible. Ceci suggère de faire le choix suivant de $t(i)$ à l'étape i de l'item $1(a)$ de l'algorithme (et c'est à cette étape que $t(i)$ doit être défini) :

on cherche $k \in \{i, \ldots, n\}$ t.q. $|a_{t(k),i}| = \max\{|a_{t(l),i}|, l \in \{i, \ldots, n\}\}$. On change alors t en prenant $t(i) = t(k)$ et $t(k) = t(i)$.

Remarque : L'introduction des matrices L et U et des vecteurs $\boldsymbol{y}$ et $\boldsymbol{x}$ n'est pas nécessaire (tout peut s'écrire avec la matrice A et le vecteur b, que l'on modifie au cours de l'algorithme). L'introduction de L, U, x et y peut toutefois aider à comprendre la méthode. Le principe retenu est que, dans les algorithmes (Gauss ou LU), on modifie la matrice A et le second membre b (en remplaçant le système à résoudre par un système équivalent) mais on ne modifie jamais L, U, y et x (qui sont définis au cours de l'algorithme).
Remarque L'algorithme se ramene donc à résoudre $LUx = b$, en faisant $L\boldsymbol{y} = b$ et $Ux = \boldsymbol{y}$. Quand on fait $L\boldsymbol{y} = b$ les equations sont dans l'ordre $t(1), \ldots, t(n)$ (les composantes de b sont donc aussi prises dans cet ordre), mais le vecteur $\boldsymbol{y}$ est bien $(y_1, \ldots, y_n)^t$ (t signifie ici "transposé" pour obtenir un vecteur colonne). Puis, on fait $Ux = \boldsymbol{y}$, les equations sont toujours dans l'ordre $t(1), \ldots, t(n)$ mais les vecteurs x et $\boldsymbol{y}$ sont $(x_1, \ldots, x_n)^t$ et $(y_1, \ldots, y_n)^t$.

1.3.4 Méthode de Choleski

On va maintenant étudier la méthode de Choleski, qui est une méthode directe adaptée au cas où A est symétrique définie positive. On rappelle qu'une matrice $A \in \mathcal{M}_N(\mathbb{R})$ de coefficients $(a_{i,j})_{i=1,N,j=1,N}$ est symétrique si $A = A^t$, où A^t désigne la transposée de A, définie par les coefficients $(a_{j,i})_{i=1,N,j=1,N}$, et que A est définie positive si $Ax \cdot x > 0$ pour tout $x \in \mathbb{R}^N$ tel que $x \neq 0$. Dans toute la suite, $x \cdot y$ désigne le produit scalaire des deux vecteurs x et y de $\mathbb{R}^N$. On rappelle (exercice) que si A est symétrique définie positive elle est en particulier inversible.

Description de la méthode

La méthode de Choleski consiste à trouver une décomposition de A de la forme $A = LL^t$, où L est triangulaire inférieure de coefficients diagonaux strictement positifs. On résout alors le système $Ax = b$ en résolvant d'abord $Ly = b$ puis le système $L^t x = y$. Une fois la matrice A "factorisée", c'est–à–dire la décomposition LL^t obtenue (voir paragraphe suivant), on effectue les étapes de "descente" et "remontée" :

1. Etape 1 : "descente" Le système $Ly = b$ s'écrit :

$$Ly = \begin{bmatrix} \ell_{1,1} & 0 & \\ \vdots & \ddots & \vdots \\ \ell_{N,1} & \ldots & \ell_{N,N} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}.$$

Ce système s'écrit composante par composante en partant de $i = 1$.

$$\begin{aligned} \ell_{1,1}y_1 = b_1, \text{ donc} \qquad & y_1 = \frac{b_1}{\ell_{1,1}} \\ \ell_{2,1}y_1 + \ell_{2,2}y_2 = b_2, \text{ donc} \qquad & y_2 = \frac{1}{\ell_{2,2}}(b_2 - \ell_{2,1}y_1) \\ \vdots \qquad & \vdots \\ \sum_{j=1,i} \ell_{i,j}y_j = b_i, \text{ donc} \qquad & y_i = \frac{1}{\ell_{i,i}}(b_i - \sum_{j=1,i-1} \ell_{i,j}y_j) \\ \vdots \qquad & \vdots \\ \sum_{j=1,N} \ell_{N,j}y_j = b_N, \text{ donc} \qquad & y_N = \frac{1}{\ell_{N,N}}(b_N - \sum_{j=1,N-1} \ell_{N,j}y_j). \end{aligned}$$

On calcule ainsi $y_1, y_2, \ldots, y_N$.

2. Etape 2 : “remontée" On calcule maintenant x solution de $L^t x = y$.

$$L^t x = \begin{bmatrix} \ell_{1,1} & \ell_{2,1} & \dots & \ell_{N,1} \\ 0 & \ddots & & \\ \vdots & & & \vdots \\ 0 & \dots & & \ell_{N,N} \end{bmatrix} \begin{bmatrix} x_1 \\ \\ \vdots \\ \\ x_N \end{bmatrix} = \begin{bmatrix} y_1 \\ \\ \vdots \\ \\ y_N \end{bmatrix}.$$

On a donc :

$$\ell_{N,N}\, x_N = y_N \text{ donc } x_N = \frac{y_N}{\ell_{N,N}}$$

$$\ell_{N-1,N-1} x_{N-1} + \ell_{N,N-1} x_N = y_{N-1} \text{ donc } x_{N-1} = \tfrac{y_{N-1} - \ell_{N,N-1} x_N}{\ell_{N-1,N-1}}$$

$$\vdots$$

$$\sum_{j=1,N} \ell_{j,1} x_j = y_1 \text{ donc } x_1 = \frac{y_1 - \sum_{j=2,N} \ell_{j,1} x_j}{\ell_{1,1}}.$$

On calcule ainsi $x_N, x_{N-1}, \dots, x_1$.

Existence et unicité de la décomposition

On donne ici le résultat d'unicité de la décomposition LL^t d'une matrice symétrique définie positive ainsi qu'un procédé constructif de la matrice L.

Théorème 1.15 (Décomposition de Choleski) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ avec $N > 1$. On suppose que A est symétrique définie positive. Alors il existe une unique matrice $L \in \mathcal{M}_N(\mathbb{R})$, $L = (\ell_{i,j})_{i,j=1}^N$, telle que :*

1. *L est triangulaire inférieure (c'est–à–dire $\ell_{i,j} = 0$ si $j > i$),*
2. *$\ell_{i,i} > 0$, pour tout $i \in \{1, \dots, N\}$,*
3. *$A = LL^t$.*

Démonstration : On sait déjà par le théorème 1.14 page 11, qi'il existe une matrice de permutation et L triangulaire inférieure et U triangulaire supérieure $PA = LU$. Ici on va montrer que dans le cas où la matrice est symétrique, la décomposition est toujours possible sans permutation. Nous donnons ici une démonstration directe de l'existence et de l'unicité de la décomposition LL^t qui a l'avantage d'être constructive.
Existence de L : démonstration par récurrence sur N

1. Dans le cas $N = 1$, on a $A = (a_{1,1})$. Comme A est symétrique définie positive, on a $a_{1,1} > 0$. On peut donc définir $L = (\ell_{1,1})$ où $\ell_{1,1} = \sqrt{a_{1,1}}$, et on a bien $A = LL^t$.
2. On suppose que la décomposition de Choleski s'obtient pour $A \in \mathcal{M}_p(\mathbb{R})$ symétrique définie positive, pour $1 \le p \le N$ et on va démontrer que la propriété est encore vraie pour $A \in \mathcal{M}_{N+1}(\mathbb{R})$ symétrique définie positive. Soit donc $A \in \mathcal{M}_{N+1}(\mathbb{R})$ symétrique définie positive ; on peut écrire A sous la forme :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.3.9}$$

où $B \in \mathcal{M}_N(\mathbb{R})$ est symétrique, $a \in \mathbb{R}^N$ et $\alpha \in \mathbb{R}$. Montrons que B est définie positive, c.à.d. que $By \cdot y > 0$, pour tout $y \in \mathbb{R}^N$ tel que $y \neq 0$. Soit donc $y \in \mathbb{R}^N \setminus \{0\}$, et $x = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^{N+1}$. Comme A est

symétrique définie positive, on a :

$$0 < Ax \cdot x = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] \cdot \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} By \\ \hline a^t y \end{array}\right] \cdot \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] = By \cdot y$$

donc B est définie positive. Par hypothèse de récurrence, il existe une matrice $M \in \mathcal{M}_N(\mathbb{R})$ $M = (m_{i,j})_{i,j=1}^N$ telle que :

(a) $m_{i,j} = 0$ si $j > i$

(b) $m_{i,i} > 0$

(c) $B = MM^t$.

On va chercher L sous la forme :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & \lambda \end{array}\right] \tag{1.3.10}$$

avec $b \in \mathbb{R}^N$, $\lambda \in \mathbb{R}_+^*$ tels que $LL^t = A$. Pour déterminer b et λ, calculons LL^t où L est de la forme (1.3.10) et identifions avec A :

$$LL^t = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & \lambda \end{array}\right] \left[\begin{array}{c|c} M^t & b \\ \hline 0 & \lambda \end{array}\right] = \left[\begin{array}{c|c} MM^t & Mb \\ \hline b^t M^t & b^t b + \lambda^2 \end{array}\right]$$

On cherche $b \in \mathbb{R}^N$ et $\lambda \in \mathbb{R}_+^*$ tels que $LL^t = A$, et on veut donc que les égalités suivantes soient vérifiées :

$$Mb = a \text{ et } b^t b + \lambda^2 = \alpha.$$

Comme M est inversible (en effet, le déterminant de M s'écrit $det(M) = \prod_{i=1}^N m_{i,i} > 0$), la première égalité ci-dessus donne : $b = M^{-1}a$ et en remplaçant dans la deuxième égalité, on obtient : $(M^{-1}a)^t(M^{-1}a) + \lambda^2 = \alpha$, donc $a^t(M^t)^{-1}M^{-1}a + \lambda^2 = \alpha$ soit encore $a^t(MM^t)^{-1}a + \lambda^2 = \alpha$, c'est–à–dire :

$$a^t B^{-1} a + \lambda^2 = \alpha \tag{1.3.11}$$

Pour que (1.3.11) soit vérifiée, il faut que

$$\alpha - a^t B^{-1} a > 0 \tag{1.3.12}$$

Montrons que la condition (1.3.12) est effectivement vérifiée : Soit $z = \begin{pmatrix} B^{-1}a \\ -1 \end{pmatrix} \in \mathbb{R}^{N+1}$. On a $z \neq 0$ et donc $Az \cdot z > 0$ car A est symétrique définie positive. Calculons Az :

$$Az = \left(\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right) \left[\begin{array}{c} B^{-1}a \\ \hline -1 \end{array}\right] = \left[\begin{array}{c} 0 \\ \\ a^t B^{-1} a - \alpha \end{array}\right].$$

On a donc $Az \cdot z = \alpha - a^t B^{-1} a > 0$ ce qui montre que (1.3.12) est vérifiée. On peut ainsi choisir $\lambda = \sqrt{\alpha - a^t B^{-1} a}$ (> 0) de telle sorte que (1.3.11) est vérifiée. Posons :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline (M^{-1}a)^t & \lambda \end{array}\right].$$

La matrice L est bien triangulaire inférieure et vérifie $\ell_{i,i} > 0$ et $A = LL^t$.

On a terminé ainsi la partie "existence".

Unicité et calcul de L. Soit $A \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive ; on vient de montrer qu'il existe $L \in \mathcal{M}_N(\mathbb{R})$ triangulaire inférieure telle que $\ell_{i,j} = 0$ si $j > i$, $\ell_{i,i} > 0$ et $A = LL^t$. On a donc :

$$a_{i,j} = \sum_{k=1}^{N} \ell_{i,k}\ell_{j,k}, \quad \forall\, (i,j) \in \{1 \dots N\}^2. \tag{1.3.13}$$

1. Calculons la 1ère colonne de L ; pour $j = 1$, on a :

$$\begin{aligned} &a_{1,1} = \ell_{1,1}\ell_{1,1} \text{ donc } \ell_{1,1} = \sqrt{a_{1,1}} \quad (a_{1,1} > 0 \text{ car } \ell_{1,1} \text{ existe }),\\ &a_{2,1} = \ell_{2,1}\ell_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{\ell_{1,1}},\\ &a_{i,1} = \ell_{i,1}\ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{\ell_{1,1}} \ \forall i \in \{2, \dots, N\}. \end{aligned}$$

2. On suppose avoir calculé les p premières colonnes de L. On calcule la colonne $(q+1)$ en prenant $j = q+1$ dans (1.3.13)

   Pour $i = q+1$, $a_{q+1,q+1} = \sum_{k=1}^{q+1} \ell_{q+1,k}\ell_{q+1,k}$ donc

$$\ell_{q+1,q+1} = (a_{q+1,q+1} - \sum_{k=1}^{q} \ell_{q+1,k}\ell_{q+1,k})^{1/2} > 0. \tag{1.3.14}$$

   Notons que $a_{q+1,q+1} - \sum_{k=1}^{q} \ell_{q+1,k}\ell_{q+1,k} > 0$ car L existe : il est indispensable d'avoir d'abord montré l'existence de L pour pouvoir exhiber le coefficient $\ell_{q+1,q+1}$.
   On procède de la même manière pour $i = q+2, \dots, N$; on a :

$$a_{i,q+1} = \sum_{k=1}^{q+1} \ell_{i,k}\ell_{q+1,k} = \sum_{k=1}^{q} \ell_{i,k}\ell_{q+1,k} + \ell_{i,q+1}\ell_{q+1,q+1}$$

   et donc

$$\ell_{i,q+1} = \left(a_{i,q+1} - \sum_{k=1}^{q} \ell_{i,k}\ell_{q+1,k} \right) \frac{1}{\ell_{q+1,q+1}}. \tag{1.3.15}$$

   On calcule ainsi toutes les colonnes de L. On a donc montré que L est unique par un moyen constructif de calcul de L.

■

Calcul du coût de la méthode de Choleski

Calcul du coût de calcul de la matrice L Dans le procédé de calcul de L exposé ci-dessus, le nombre d'opérations pour calculer la première colonne est N. Calculons, pour $p = 0, \dots, N-1$, le nombre d'opérations pour calculer la $(p+1)$-ième colonne : pour la colonne $(p+1)$, le nombre d'opérations par ligne est $2p+1$, car le calcul de $\ell_{p+1,p+1}$ par la formule (1.3.14) nécessite p multiplications, p soustractions et une extraction de racine, soit $2p+1$ opérations ; le calcul de $\ell_{i,p+1}$ par la formule (1.3.15) nécessite p multiplications, p soustractions et

une division, soit encore $2p+1$ opérations. Comme les calculs se font des lignes $p+1$ à N (car $\ell_{i,p+1}=0$ pour $i \leq p$), le nombre d'opérations pour calculer la $(p+1)$-ième colonne est donc $(2p+1)(N-p)$. On en déduit que le nombre d'opérations N_L nécessaires au calcul de L est :

$$\begin{aligned} N_L &= \sum_{p=0}^{N-1}(2p+1)(N-p) = 2N\sum_{p=0}^{N-1}p - 2\sum_{p=0}^{N-1}p^2 + N\sum_{p=0}^{N-1}1 - \sum_{p=0}^{N-1}p \\ &= (2N-1)\frac{N(N-1)}{2} + N^2 - 2\sum_{p=0}^{N-1}p^2. \end{aligned}$$

(On rappelle que $2\sum_{p=0}^{N-1} p = N(N-1)$.) Il reste à calculer $C_N = \sum_{p=0}^{N} p^2$, en remarquant par exemple que

$$\begin{aligned} \sum_{p=0}^{N}(1+p)^3 &= \sum_{p=0}^{N} 1+p^3+3p^2+3p = \sum_{p=0}^{N}1 + \sum_{p=0}^{N}p^3 + 3\sum_{p=0}^{N}p^2 + 3\sum_{p=0}^{N}p \\ &= \sum_{p=1}^{N+1}p^3 = \sum_{p=0}^{N}p^3 + (N+1)^3. \end{aligned}$$

On a donc $3C_N + 3\frac{N(N+1)}{2} + N + 1 = (N+1)^3$, d'où on déduit que

$$C_N = \frac{N(N+1)(2N+1)}{6}.$$

On a donc :

$$\begin{aligned} N_L &= (2N-1)\frac{N(N-1)}{2} - 2C_{N-1} + N^2 \\ &= N\left(\frac{2N^2+3N+1}{6}\right) = \frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6} = \frac{N^3}{3} + 0(N^2). \end{aligned}$$

Coût de la résolution d'un système linéaire par la méthode LL^t Nous pouvons maintenant calculer le coût (en termes de nombre d'opérations élémentaires) nécessaire à la résolution de (1.1.1) par la méthode de Choleski pour $A \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive. On a besoin de N_L opérations pour le calcul de L, auquel il faut rajouter le nombre d'opérations nécessaires pour les étapes de descente et remontée. Le calcul de y solution de $Ly = b$ s'effectue en résolvant le système :

$$\begin{bmatrix} \ell_{1,1} & & 0 \\ \vdots & \ddots & \vdots \\ \ell_{N,1} & \dots & \ell_{N,1} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}$$

Pour la ligne 1, le calcul $y_1 = \dfrac{b_1}{\ell_{1,1}}$ s'effectue en une opération.

Pour les lignes $p = 2$ à N, le calcul $y_p = \left(b_p - \sum_{i=1}^{p-1}\ell_{i,p}y_i\right)/\ell_{p,p}$ s'effectue en $(p-1)$ (multiplications) $+(p-2)$ (additions) $+1$ soustraction $+1$ (division) $= 2p-1$ opérations. Le calcul de y (descente) s'effectue donc en $N_1 = \sum_{p=1}^{N}(2p-1) = N(N+1) - N = N^2$. On peut calculer de manière similaire le nombre d'opérations nécessaires pour l'étape de remontée $N_2 = N^2$. Le nombre total d'opérations pour calculer x solution de (1.1.1) par la méthode de Choleski est $N_C = N_L + N_1 + N_2 = \frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6} + 2N^2 = \frac{N^3}{3} + \frac{5N^2}{2} + \frac{N}{6}$. L'étape la plus coûteuse est donc la factorisation de A.

Remarque 1.16 (Décomposition LDL^t) *Dans les programmes informatiques, on préfère implanter la variante suivante de la décomposition de Choleski : $A = \tilde{L}D\tilde{L}^t$ où D est la matrice diagonale définie par $d_{i,i} = \ell_{i,i}^2$, $\tilde{L}_{i,i} = L\tilde{D}^{-1}$, où $\tilde{D}$ est la matrice diagonale définie par $d_{i,i} = \ell_{i,i}$. Cette décomposition a l'avantage de ne pas faire intervenir le calcul de racines carrées, qui est une opération plus compliquée que les opérations "élémentaires" $(\times, +, -)$.*

1.3.5 Quelques propriétés

Comparaison Gauss/Choleski

Soit $A \in \mathcal{M}_N(\mathbb{R})$ inversible, la résolution de (1.1.1) par la méthode de Gauss demande $2N^3/3+0(N^2)$ opérations (exercice). Dans le cas d'une matrice symétrique définie positive, la méthode de Choleski est donc environ deux fois moins chère.

Et la méthode de Cramer ?

Soit $A \in \mathcal{M}_N(\mathbb{R})$ inversible. On rappelle que la méthode de Cramer pour la résolution de (1.1.1) consiste à calculer les composantes de x par les formules :

$$x_i = \frac{det(A_i)}{det(A)},\ i = 1, \ldots, N,$$

où A_i est la matrice carrée d'ordre N obtenue à partir de A en remplaçant la i-ème colonne de A par le vecteur b, et $det(A)$ désigne le déterminant de A.
Chaque calcul de déterminant d'une matrice carrée d'ordre N nécessite au moins $N!$ opérations (voir cours L1-L2, ou livres d'algèbre linéaire proposés en avant-propos). Par exemple, pour $N = 10$, la méthode de Gauss nécessite environ 700 opérations, la méthode de Choleski environ 350 et la méthode de Cramer plus de 4 000 000.... Cette dernière méthode est donc à proscrire.

Conservation du profil de A

Dans de nombreuses applications, par exemple lors de la résolution de systèmes linéaires issus de la discrétisation [1] de problèmes réels, la matrice $A \in \mathcal{M}_N(\mathbb{R})$ est "creuse", au sens où un grand nombre de ses coefficients sont nuls. Il est intéressant dans ce cas pour des raisons d'économie de mémoire de connaître le "profil" de la matrice, donné dans le cas où la matrice est symétrique, par les indices $j_i = \min\{j \in \{1, \ldots, N\}$ tels que $a_{i,j} \neq 0\}$. Le profil de la matrice est donc déterminé par les diagonales contenant des coefficients non nuls qui sont les plus éloignées de la diagonale principale. Dans le cas d'une matrice creuse, il est avantageux de faire un stockage "profil" de A, en stockant, pour chaque ligne i la valeur de j_i et des coefficients $a_{i,k}$, pour $k = i - j_i, \ldots, i$, ce qui peut permettre un large gain de place mémoire, comme on peut s'en rendre compte sur la figure 1.3.5.
Une propriété intéressante de la méthode de Choleski est de conserver le profil. On peut montrer (en reprenant les calculs effectués dans la deuxième partie de la démonstration du théorème 1.15) que $\ell_{i,j} = 0$ si $j < j_i$. Donc si on a adopté un stockage "profil" de A, on peut utiliser le même stockage pour L.

Matrices non symétriques

Soit $A \in \mathcal{M}_N(\mathbb{R})$ inversible. On ne suppose plus ici que A est symétrique. On cherche à calculer $x \in \mathbb{R}^N$ solution de (1.1.1) par la méthode de Choleski. Ceci est possible en remarquant que : $Ax = b \Leftrightarrow A^tAx = A^tb$ car $det(A) = det(A^t) \neq 0$. Il ne reste alors plus qu'à vérifier que A^tA est symétrique définie positive. Remarquons d'abord que pour toute matrice $A \in \mathcal{M}_N(\mathbb{R})$, la matrice AA^t est symétrique. Pour cela on utilise le fait que

1. On appelle discrétisation le fait de se ramer d'un problème où l'inconnue est une fonction en un problème ayant un nombre fini d'inconnues.

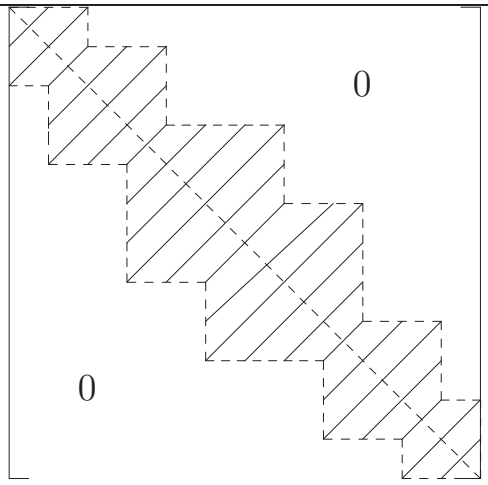

FIGURE 1.3 – Exemple de profil d'une matrice symétrique

si $B \in \mathcal{M}_N(\mathbb{R})$, alors B est symétrique si et seulement si $Bx \cdot y = x \cdot By$ et $Bx \cdot y = x \cdot B^t y$ pour tout $(x, y) \in (\mathbb{R}^N)^2$. En prenant $B = A^t A$, on en déduit que $A^t A$ est symétrique. De plus, comme A est inversible, $A^t Ax \cdot x = Ax \cdot Ax = |Ax|^2 > 0$ si $x \neq 0$. La matrice $A^t A$ est donc bien symétrique définie positive.
La méthode de Choleski dans le cas d'une matrice non symétrique consiste donc à calculer $A^t A$ et $A^t b$, puis à résoudre le système linéaire$A^t A \cdot x = A^t b$ par la méthode de Choleski "symétrique".
Cette manière de faire est plutôt moins efficace que la décomposition LU puisque le coût de la décomposition LU est de $2N^3/3$ alors que la méthode de Choleski dans le cas d'une matrice non symétrique nécessite au moins $4N^3/3$ opérations (voir exercice 14).

Systèmes linéaires non carrés

On considère ici des matrices qui ne sont plus carrées. On désigne par $\mathcal{M}_{M,N}(\mathbb{R})$ l'ensemble des matrices réelles à M lignes et N colonnes. Pour $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $M > N$ et $b \in \mathbb{R}^M$, on cherche $x \in \mathbb{R}^N$ tel que

$$Ax = b. \tag{1.3.16}$$

Ce système contient plus d'équations que d'inconnues et n'admet donc en général pas de solution. On cherche $x \in \mathbb{R}^N$ qui vérifie le sytème (1.3.16) "au mieux". On introduit pour cela une fonction f définie de$\mathbb{R}^N$ dans $\mathbb{R}$ par :

$$f(x) = |Ax - b|^2,$$

où $|x| = \sqrt{x \cdot x}$ désigne la norme euclidienne sur $\mathbb{R}^N$. La fonction f ainsi définie est évidemment positive, et s'il existe x qui annule f, alors x est solution du système (1.3.16). Comme on l'a dit, un tel x n'existe pas forcément, et on cherche alors un vecteur x qui vérifie (1.3.16) "au mieux", au sens où $f(x)$ soit le plus proche de 0. On cherche donc $x \in \mathbb{R}^N$ satisfaisant (1.3.16) en minimisant f, c.à.d. en cherchant $x \in \mathbb{R}^N$ solution du problème d'optimisation :

$$f(x) \leq f(y) \quad \forall y \in \mathbb{R}^N \tag{1.3.17}$$

On peut réécrire f sous la forme : $f(x) = A^t Ax \cdot x - 2b \cdot Ax + b \cdot b$. On montrera au chapitre III que s'il existe une solution au problème (1.3.17), elle est donnée par la résolution du système linéaire suivant :

$$AA^t x = A^t b \ \in \mathbb{R}^N,$$

qu'on appelle équations normales du problème de minimisation. La résolution approchée du problème (1.3.16) par cette procédure est appelée méthode des moindres carrés . La matrice AA^t étant symétrique, on peut alors employer la méthode de Choleski pour la résolution du système (1.3.5).

1.4 Conditionnement

Dans ce paragraphe, nous allons définir la notion de conditionnement d'une matrice, qui peut servir à établir une majoration des erreurs d'arrondi dues aux erreurs sur les données. Malheureusement, nous verrons également que cette majoration n'est pas forcément très utile dans des cas pratiques, et nous nous efforcerons d'y remédier. la notion de conditionnement est également utilisée dans l'étude des méthodes itératives que nous verrons plus loin.

1.4.1 Le problème des erreurs d'arrondis

Soient $A \in \mathcal{M}_N(\mathbb{R})$ inversible et $b \in \mathbb{R}^N$; supposons que les données A et b ne soient connues qu'à une erreur près. Ceci est souvent le cas dans les applications pratiques. Considérons par exemple le problème de la conduction thermique dans une tige métallique de longueur 1, modélisée par l'intervalle $[0, 1]$. Supposons que la température u de la tige soit imposée aux extrémités, $u(0) = u_0$ et $u(1) = u_1$. On suppose que la température dans la tige satisfait à l'équation de conduction de la chaleur, qui s'écrit $(k(x)u'(x))' = 0$, où k est la conductivité thermique. Cette équation différentielle du second ordre peut se discrétiser par exemple par différences finies (on verra une description de la méthode page 23), et donne lieu à un système linéaire de matrice A. Si la conductivité k n'est connue qu'avec une certaine précision, alors la matrice A sera également connue à une erreur près, notée δ_A. On aimerait que l'erreur commise sur les données du modèle (ici la conductivité thermique k) n'ait pas une conséquence catastrophique sur le calcul de la solution du modèle (ici la température u). Si par exemple 1% d'erreur sur k entraîne 100% d'erreur sur u, le modèle ne sera pas d'une utilité redoutable...
L'objectif est donc d'estimer les erreurs commises sur x solution de (1.1.1) à partir des erreurs commises sur b et A. Notons $\delta_b \in \mathbb{R}^N$ l'erreur commise sur b et $\delta_A \in \mathcal{M}_N(\mathbb{R})$ l'erreur commise sur A. On cherche alors à évaluer δ_x où $x + \delta_x$ est solution (si elle existe) du système :

$$\begin{cases} x + \delta_x \in \mathbb{R}^N \\ (A + \delta_A)(x + \delta_x) = b + \delta_b. \end{cases} \tag{1.4.18}$$

On va montrer que si δ_A "n'est pas trop grand", alors la matrice $A + \delta_A$ est inversible, et qu'on peut estimer δ_x en fonction de δ_A et δ_b.

1.4.2 Conditionnement et majoration de l'erreur d'arrondi

Définition 1.17 (Conditionnement) *Soit $\mathbb{R}^N$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_N(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On appelle conditionnement de A par rapport à la norme $\|\cdot\|$ le nombre réel positif $\mathrm{cond}(A)$ défini par :*

$$\mathrm{cond}(A) = \|A\| \ \|A^{-1}\|.$$

Proposition 1.18 (Propriétés générales du conditionnement) *Soit $\mathbb{R}^N$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_N(\mathbb{R})$ muni de la norme induite.*

1. *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, alors $\mathrm{cond}(A) \geq 1$.*
2. *Soit $A \in \mathcal{M}_N(\mathbb{R})$ et $\alpha \in \mathbb{R}^*$, alors $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$.*
3. *Soient A et $B \in \mathcal{M}_N(\mathbb{R})$ des matrices inversibles, alors $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.*

Proposition 1.19 (Propriétés du conditionnement pour la norme 2) *Soit $\mathbb{R}^N$ muni de la norme euclidienne $\|\cdot\|_2$ et $\mathcal{M}_N(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^N$.*

1. *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On note σ_N [resp. σ_1] la plus grande [resp. petite] valeur propre de A^tA (noter que A^tA est une matrice symétrique définie positive). Alors* $\mathrm{cond}_2(A) = \sqrt{\sigma_N/\sigma_1}$.
2. *Si de plus A une matrice symétrique définie positive, alors* $\mathrm{cond}_2(A) = \frac{\lambda_N}{\lambda_1}$, *où λ_N [resp. λ_1] est la plus grande [resp. petite] valeur propre de A.*
3. *Si A et B sont deux matrices symétriques définies positives, alors* $\mathrm{cond}_2(A+B) \leq \max(\mathrm{cond}_2(A), \mathrm{cond}_2(B))$.
4. *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. Alors* $\mathrm{cond}_2(A) = 1$ *si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^\star$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).*
5. *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Alors* $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.
6. *Soient A, $B \in \mathcal{M}_N(\mathbb{R})$ deux matrices symétriques définies positives. Montrer que* $\mathrm{cond}_2(A+B) \leq \max\{\mathrm{cond}_2(A), \mathrm{cond}_2(B)\}$.

La démonstration de deux propositions précédentes fait l'objet de l'exercice 18 page 40.

Théorème 1.20 *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^N$, $b \neq 0$. On munit $\mathbb{R}^N$ d'une norme $\|\cdot\|$, et $\mathcal{M}_N(\mathbb{R})$ de la norme induite. Soient $\delta_A \in \mathcal{M}_N(\mathbb{R})$ et $\delta_b \in \mathbb{R}^N$. On suppose que $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$. Alors la matrice $(A+\delta_A)$ est inversible et si x est solution de* (1.1.1) *et $x+\delta_x$ est solution de* (1.4.18)*, alors*

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\delta_b\|}{\|b\|} + \frac{\|\delta_A\|}{\|A\|}\right). \tag{1.4.19}$$

Démonstration :
On peut écrire $A + \delta_A = A(Id + B)$ avec $B = A^{-1}\delta_A$. Or le rayon spectral de B, $\rho(B)$, vérifie $\rho(B) \leq \|B\| \leq \|\delta_A\|\ \|A^{-1}\| < 1$, et donc (voir le théorème 1.9 page 7 et l'exercice 9 page 37) $(Id + B)$ est inversible et $(Id+B)^{-1} = \sum_{n=0}^{\infty}(-1)^n B^n$. On a aussi $\|(Id+B)^{-1}\| \leq \sum_{n=0}^{\infty}\|B\|^n = \dfrac{1}{1-\|B\|} \leq \dfrac{1}{1-\|A^{-1}\|\ \|\delta_A\|}$. On en déduit que $A+\delta_A$ est inversible, car $A+\delta_A = A(Id+B)$ et comme A est inversible, $(A+\delta_A)^{-1} = (Id+B)^{-1}A^{-1}$.

Comme A et $A+\delta_A$ sont inversibles, il existe un unique $x \in \mathbb{R}^N$ tel que $Ax = b$ et il existe un unique $\delta_x \in \mathbb{R}^N$ tel que $(A+\delta_A)(x+\delta_x) = b+\delta_b$. Comme $Ax = b$, on a $(A+\delta_A)\delta_x + \delta_A x = \delta_b$ et donc $\delta_x = (A+\delta_A)^{-1}(\delta_b - \delta_A x)$. Or $(A+\delta_A)^{-1} = (Id+B)^{-1}A^{-1}$, on en déduit :

$$\begin{aligned}\|(A+\delta_A)^{-1}\| &\leq \|(Id+B)^{-1}\|\ \|A^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\ \|\delta_A\|}.\end{aligned}$$

On peut donc écrire la majoration suivante :

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\|A^{-1}\|\ \|A\|}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\delta_b\|}{\|A\|\ \|x\|} + \frac{\|\delta_A\|}{\|A\|}\right).$$

En utilisant le fait que $b = Ax$ et que par suite $\|b\| \leq \|A\|\ \|x\|$, on obtient :

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\|A^{-1}\|\ \|A\|}{1-\|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta_A\|}{\|A\|}\right),$$

ce qui termine la démonstration. ■

Remarquons que l'estimation (1.4.19) est optimale. En effet, en supposant que $\delta_A = 0$. L'estimation (1.4.19) devient alors :

$$\frac{\|\delta_x\|}{\|x\|} \leq cond(A)\frac{\|\delta_b\|}{\|b\|}. \tag{1.4.20}$$

Peut-on avoir égalité dans (1.4.20) ? Pour avoir égalité dans (1.4.20) il faut choisir convenablement b et δ_b. Soit $x \in \mathbb{R}^N$ tel que $\|x\| = 1$ et $\|Ax\| = \|A\|$. Notons qu'un tel x existe parce que $\|A\| = \sup\{\|Ax\|; \|x\| = 1\} = \max\{\|Ax\|; \|x\| = 1\}$ (voir proposition 1.2 page 5) Posons $b = Ax$; on a donc $\|b\| = \|A\|$. De même, grâce à la proposition 1.2, il existe $y \in \mathbb{R}^N$ tel que $\|y\| = 1$, et $\|A^{-1}y\| = \|A^{-1}\|$. On choisit alors δ_b tel que $\delta_b = \varepsilon y$ où $\varepsilon > 0$ est donné. Comme $A(x + \delta_x) = b + \delta_b$, on a $\delta_x = A^{-1}\delta_b$ et donc : $\|\delta_x\| = \|A^{-1}\delta_b\| = \varepsilon\|A^{-1}y\| = \varepsilon\|A^{-1}\| = \|\delta_b\| \ \|A^{-1}\|$. On en déduit que

$$\frac{\|\delta_x\|}{\|x\|} = \|\delta_x\| = \|\delta_b\| \ \|A^{-1}\| \ \ \frac{\|A\|}{\|b\|} \text{ car } \|b\| = \|A\| \text{ et } \|x\| = 1.$$

Par ce choix de b et δ_b on a bien égalité dans (1.4.20). L'estimation (1.4.20) est donc optimale.

1.4.3 Discrétisation d'équations différentielles, conditionnement "efficace"

On suppose encore ici que $\delta_A = 0$. On suppose que la matrice A du système linéaire à résoudre provient de la discrétisation par différences finies d'une équation différentielle introduite ci-dessous (voir (1.4.21)). On peut alors montrer (voir exercice 26 page 44 du chapitre 1) que le conditionnement de A est d'ordre N^2, où N est le nombre de points de discrétisation. Pour $N = 10$, on a donc $\mathrm{cond}(A) \simeq 100$ et l'estimation (1.4.20) donne :

$$\frac{\|\delta_x\|}{\|x\|} \leq 100 \ \frac{\|\delta_b\|}{\|b\|}.$$

Une erreur de 1% sur b peut donc entraîner une erreur de 100% sur x. Autant dire que dans ce cas, il est inutile de rechercher la solution de l'équation discrétisée...Heureusement, on peut montrer que l'estimation (1.4.20) n'est pas significative pour l'étude de la propagation des erreurs lors de larésolution des systèmes linéraires provenant de la discrétisation d'une équation différentielle ou d'une équation aux dérivées partielles [2]. Pour illustrer notre propos, nous allons étudier un système linéaire très simple provenant d'un problème de mécanique et sa discrétisation par différences finies.

Discrétisation par différences finies de $-u'' = f$ Soit $f \in C([0,1], \mathbb{R})$. On cherche u tel que

$$\begin{cases} -u''(x) = f(x) \\ u(0) = u(1) = 0. \end{cases} \tag{1.4.21}$$

On peut montrer (on l'admettra ici) qu'il existe une unique solution $u \in C^2([0,1], \mathbb{R})$. On cherche à calculer u de manière approchée. On va pour cela introduire la méthode de discrétisation dite *par différences finies*. Soit $N \in \mathbb{N}^*$, on définit $h = 1/(N+1)$ le *pas de discrétisation*, c.à.d. la distance entre deux points de discrétisation, et pour $i = 0 \ldots N+1$ on définit les points de discrétisation $x_i = ih$ (voir Figure 1.4.3), qui sont les points où l'on va écrire l'équation $-u'' = f$ en vue de se ramer à un système discret, c.à.d. à un système avec un nombre fini d'inconnues. Remarquons que $x_0 = 0$ et $x_{N+1} = 1$. Soit $u(x_i)$ la valeur exacte de u en x_i. On écrit la première équation de (1.4.21) en chaque point x_i, pour $i = 1 \ldots N$.

$$-u''(x_i) = f(x_i) = b_i \ \forall i \in \{1 \ldots N\}.$$

2. On appelle équation aux dérivées partielles une équation qui fait intervenir les dérivées partielles de la fonction inconnue, par exemple $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$, où u est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$

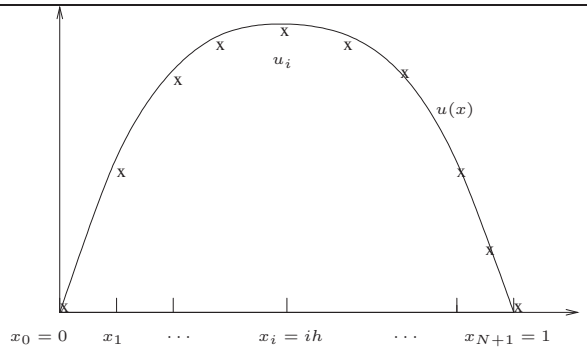

FIGURE 1.4 – Solution exacte et approchée de $-u'' = f$

On peut facilement montrer, par un développement de Taylor, que si $u \in C^4([0,1], \mathbb{R})$ (ce qui est vrai si $f \in C^2$) alors [3]

$$-\frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} = -u''(x_i) + R_i \text{ avec } |R_i| \leq \frac{h^2}{12}\|u^{(4)}\|_\infty. \tag{1.4.22}$$

La valeur R_i s'appelle erreur de consistance au point x_i.
On introduit alors les inconnues $(u_i)_{i=1,\ldots,N}$ qu'on espère être des valeurs approchées de u aux points x_i et qui sont les composantes de la solution (si elle existe) du système suivant

$$\begin{cases} -\dfrac{u_{i+1} + u_{i-1} - 2u_i}{h^2} = b_i, \quad \forall i \in \{1 \leq N\}, \\ u_0 = u_{N+1} = 0. \end{cases} \tag{1.4.23}$$

On cherche donc $u = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix} \in \mathbb{R}^N$ solution de (1.4.23). Ce système peut s'écrire sous forme matricielle :$Au = b$ avec$b = (b_1, \ldots, b_N)^t$ et A la matrice carrée d'ordre N de coefficients $(a_{i,j})_{i,j=1,N}$ définis par :

$$\begin{cases} a_{i,i} &= \dfrac{2}{h^2}, \forall\, i = 1, \ldots, N, \\ a_{i,j} &= -\dfrac{1}{h^2}, \forall\, i = 1, \ldots, N, \ \ j = i \pm 1, \\ a_{i,j} &= 0, \forall\, i = 1, \ldots, N, \ |i - j| > 1. \end{cases} \tag{1.4.24}$$

On remarque immédiatement que A est tridiagonale. On peut montrer que A est symétrique définie positive (voir exercice 26 page 44). On peut aussi montrer que

$$\max_{i=1\ldots N}\{|u_i - u(x_i)|\} \leq \frac{h^2}{96}\|u^{(4)}\|_\infty.$$

En effet, si on note $\overline{u}$ le vecteur de $\mathbb{R}^N$ de composantes $u(x_i)$, $i = 1, \ldots, N$, et R le vecteur de $\mathbb{R}^N$ de composantes R_i, $i = 1, \ldots, N$, on a par définition de R (formule (1.4.22)) $A(u - \overline{u}) = R$, et donc $\|u - \overline{u}\|_\infty \leq$

3. En effet, par développement de Taylor, on a :
$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i)$ où $\xi_i \in (x_i, x_{i+1})$ et
$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i)$ où $\eta_i \in (x_i, x_{i+1})$.
En faisant la somme de ces deux égalités, on en déduit que : $u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + \frac{h^2}{u}''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i) + \frac{h^4}{24}u^{(4)}(\eta_i)$, et donc R_i défini par (1.4.22) vérifie :
$|R_i| = |-\frac{u(x_{i+1})+u(x_{i-1})-2u(x_i)}{h^2} + u''(x_i)| \leq |\frac{h^4}{24}u^{(4)}(\xi_i) + \frac{h^4}{24}u^{(4)}(\eta_i)| \leq \frac{h^2}{12}\|u^{(4)}\|_\infty$.

$\|A^{-1}\|_\infty \|R\|_\infty$. Or on peut montrer (voir exercice 28 page 45, partie I) que $\|A^{-1}\|_\infty \leq 1/8$, et on obtient donc avec (1.4.22) que $\|u - \overline{u}\|_\infty \leq h^2/96 \|u^{(4)}\|_\infty$.

Cette inégalité donne la précision de la méthode. On remarque en particulier que si on raffine la discrétisation, c'est–à–dire si on augmente le nombre de points N ou, ce qui revient au même, si on diminue le pas de discrétisation h, on augmente la précision avec laquelle on calcule la solution approchée. Or on a déjà dit qu'on peut montrer (voir exercice 26 page 44) que $\mathrm{cond}(A) \simeq N^2$. Donc si on augmente le nombre de points, le conditionnement de A augmente aussi. Par exemple si $N = 10^4$, alors $\|\delta_x\|/\|x\| = 10^8 \|\delta_b\|/\|b\|$. Or sur un ordinateur en simple précision, on a $\|\delta_b\|/\|b\| \geq 10^{-7}$, donc l'estimation (1.4.20) donne une estimation de l'erreur relative $\|\delta_x\|/\|x\|$ de 1000%, ce qui laisse à désirer pour un calcul qu'on espère précis.
En fait, l'estimation (1.4.20) ne sert à rien pour ce genre de problème, il faut faire une analyse un peu plus poussée, comme c'est fait dans l'exercice 28 page 45. On se rend compte alors que pour f donnée il existe $C \in \mathbb{R}_+$ ne dépendant que de f (mais pas de N) tel que

$$\frac{\|\delta_u\|}{\|u\|} \leq C \frac{\|\delta_b\|}{\|b\|} \text{ avec } b = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix}. \tag{1.4.25}$$

L'estimation (1.4.25) est évidemment bien meilleure que l'estimation (1.4.20) puisqu'elle montre que l'erreur relative commise sur u est du même ordre que celle commise sur b. En particulier, elle n'augmente pas avec le nombre de points de discrétisation. En conclusion, l'estimation (1.4.20) est peut-être optimale dans le cas d'une matrice quelconque, (on a montré ci-dessus qu'il peut y avoir égalité dans (1.4.20)) mais elle n'est pas significative pour l'étude des systèmes linéaires issus de la discrétisation des équations aux dérivées partielles.

1.5 Méthodes itératives

1.5.1 Origine des systèmes à résoudre

Les méthodes directes que nous avons étudiées dans le paragraphe précédent sont très efficaces : elles donnent la solution exacte (aux erreurs d'arrondi près) du système linéaire considéré. Elles ont l'inconvénient de nécessiter une assez grande place mémoire car elles nécessitent le stockage de toute la matrice en mémoire vive. Si la matrice est pleine, c.à.d. si la plupart des coefficients de la matrice sont non nuls et qu'elle est trop grosse pour la mémoire vive de l'ordinateur dont on dispose, il ne reste plus qu'à gérer habilement le "swapping" c'est–à–dire l'échange de données entre mémoire disque et mémoire vive pour pouvoir résoudre le système.
Cependant, si le système a été obtenu à partir de la discrétisation d'équations aux dérivés partielles, il est en général "creux", c.à. d. qu'un grand nombre des coefficients de la matrice du système sont nuls ; de plus la matrice a souvent une structure "bande", *i.e.* les éléments non nuls de la matrice sont localisés sur certaines diagonales. On a vu au chapitre précédent que dans ce cas, la méthode de Choleski "conserve le profil" (voir à ce propos page 19). Prenons par exemple le cas d'une discrétisation du Laplacien sur un carré par différences finies. On cherche à résoudre le problème :

$$\begin{array}{l} -\Delta u = f \text{ sur } \Omega =]0,1[\times]0,1[, \\ u = 0 \text{ sur } \partial\Omega, \end{array} \tag{1.5.26}$$

On rappelle que l'opérateur Laplacien est défini pour $u \in C^2(\Omega)$, où Ω est un ouvert de $\mathbb{R}^2$, par

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Définissons une discrétisation uniforme du carré par les points (x_i, y_j), pour $i = 1, \ldots, M$ et $j = 1, \ldots, M$ avec $x_i = ih$, $y_j = jh$ et $h = 1/(M+1)$, representée en figure 1.5.1 pour $M = 6$. On peut alors approcher les

dérivées secondes par des quotients différentiels comme dans le cas unidimensionnel (voir page 23), pour obtenir un système linéaire : $AU = b$ où $A \in \mathcal{M}_N(\mathbb{R})$ et $b \in \mathbb{R}^N$ avec $N = M^2$. Utilisons l'ordre"*lexicographique*" pour numéroter les inconnues, c.à.d. de bas en haut et de gauche à droite : les inconnues sont alors numérotées de 1 à $N = M^2$ et le second membre s'écrit $b = (b_1, \ldots, b_N)^t$. Les composantes $b_1, \ldots, b_N$ sont définies par :pour $i, j = 1, \ldots, M$, on pose $k = j + (i-1)M$ et $b_k = f(x_i, y_j)$.

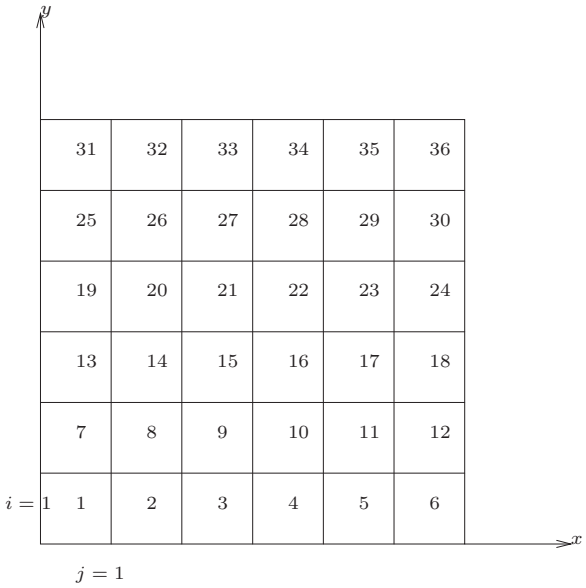

FIGURE 1.5 – Ordre lexicographique des inconnues, exemple dans le cas $M = 6$

Les coefficients de $A = (a_{k,\ell})_{k,l=1,N}$ peuvent être calculés de la manière suivante :

$$
\left\{
\begin{array}{ll}
\text{Pour} & i, j = 1, \ldots, M, \text{ on pose } k = j + (i-1)M, \\
a_{k,k} & = \dfrac{4}{h^2}, \\
a_{k,k+1} = & \left\{ \begin{array}{ll} -\dfrac{1}{h^2} & \text{si } j \neq M, \\ 0 & \text{sinon,} \end{array} \right. \\
a_{k,k-1} = & \left\{ \begin{array}{ll} -\dfrac{1}{h^2} & \text{si } j \neq 1, \\ 0 & \text{sinon,} \end{array} \right. \\
a_{k,k+M} = & \left\{ \begin{array}{ll} -\dfrac{1}{h^2} & \text{si } i < M, \\ 0 & \text{sinon,} \end{array} \right. \\
a_{k,k-M} = & \left\{ \begin{array}{ll} -\dfrac{1}{h^2} & \text{si } i > 1, \\ 0 & \text{sinon,} \end{array} \right. \\
\text{Pour} & k = 1, \ldots, N, \text{ et } \ell = 1, \ldots, N; \\
a_{k,\ell} & = 0, \forall\, k = 1, \ldots, N,\ 1 < |k - \ell| < N \text{ ou } |k - \ell| > N.
\end{array}
\right.
$$

La matrice est donc tridiagonale par blocs, plus précisément si on note

$$D = \begin{pmatrix} 4 & -1 & 0 & \dots & \dots & 0 \\ -1 & 4 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \\ 0 & & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & & 0 & -1 & 4 \end{pmatrix},$$

les blocs diagonaux (qui sont des matrices de dimension $M \times M$), on a :

$$A = \begin{pmatrix} D & -Id & 0 & \dots & \dots & 0 \\ -Id & D & -Id & 0 & \dots & 0 \\ 0 & -Id & D & -Id & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & -Id & D & -Id \\ 0 & \dots & & 0 & -Id & D \end{pmatrix}, \tag{1.5.27}$$

où Id désigne la matrice identité d'ordre M.
On peut remarquer que la matrice A a une "largeur de bande" de M. Si on utilise une méthode directe genre Choleski, on aura donc besoin d'une place mémoire de $N \times M = M^3$. (Notons que pour une matrice pleine on a besoin de M^4.)
Lorsqu'on a affaire à de très gros systèmes issus par exemple de l'ingénierie (calcul des structures, mécanique des fluides, ...), où N peut être de l'ordre de plusieurs milliers, on cherche à utiliser des méthodes nécessitant le moins de mémoire possible. On a intérêt dans ce cas à utiliser des méthodes itératives. Ces méthodes ne font appel qu'à des produits matrice vecteur, et ne nécessitent donc pas le stockage du profil de la matrice mais uniquement des termes non nuls. Dans l'exemple précédent, on a 5 diagonales non nulles, donc la place mémoire nécessaire pour un produit matrice vecteur est $5N = 5M^2$. Ainsi pour les gros systèmes, il est souvent avantageux d'utiliser des méthodes itératives qui ne donnent pas toujours la solution exacte du système en un nombre fini d'itérations, mais qui donnent une solution approchée à coût moindre qu'une méthode directe, car elles ne font appel qu'à des produits matrice vecteur.

Remarque 1.21 (Sur la méthode du gradient conjugué)
Il existe une méthode itérative "miraculeuse" de résolution des systèmes linéaires lorsque la matrice A est symétrique définie positive : c'est la méthode du gradient conjugué. Elle est miraculeuse en ce sens qu'elle donne la solution exacte du système $Ax = b$ en un nombre fini d'opérations (en ce sens c'est une méthode directe) : moins de N itérations où N est l'ordre de la matrice A, bien qu'elle ne nécessite que des produits matrice vecteur ou des produits scalaires. La méthode du gradient conjugué est en fait une méthode d'optimisation pour la recherche du minimum dans $\mathbb{R}^N$ de la fonction de $\mathbb{R}^N$ dans $\mathbb{R}$ définie par : $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Or on peut montrer que lorsque A est symétrique définie positive, la recherche de x minimisant f dans $\mathbb{R}^N$ est équivalent à la résolution du système $Ax = b$. (Voir paragraphe 3.2.2 page 128.) En fait, la méthode du gradient conjugué n'est pas si miraculeuse que cela en pratique : en effet, le nombre N est en général très grand et on ne peut en général pas envisager d'effectuer un tel nombre d'itérations pour résoudre le système. De plus, si on utilise la méthode du gradient conjugué brutalement, non seulement elle ne donne pas la solution en N itérations en raison de l'accumulation des erreurs d'arrondi, mais plus la taille du système croît et plus le nombre d'itérations nécessaires devient élevé. On a alors recours aux techniques de "préconditionnement". Nous reviendrons sur ce point au chapitre 3.
La méthode itérative du gradient à pas fixe, qui est elle aussi obtenue comme méthode de minimisation de la fonction f ci-dessus, fait l'objet des exercices 29 page 46 et 68 page 153.

1.5.2 Définition et propriétés

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^N$, on cherche toujours ici à résoudre le système linéaire (1.1.1) c'est–à–dire à trouver $x \in \mathbb{R}^N$ tel que $Ax = b$.

Définition 1.22 *On appelle méthode itérative de résolution du système linéaire* (1.1.1) *une méthode qui construit une suite* $(x^{(k)})_{k\in\mathbb{N}}$ *(où "l'itéré" $x^{(k)}$ est calculé à partir des itérés $x^{(0)} \dots x^{(k-1)}$) censée converger vers x solution de* (1.1.1)*.*

Définition 1.23 *On dit qu'une méthode itérative est convergente si pour tout choix initial $x^{(0)} \in \mathbb{R}^N$, on a :*

$$x^{(k)} \longrightarrow x \text{ quand } n \to +\infty$$

Puisqu'il s'agit de résoudre un sytème linéaire, il est naturel d'essayer de construire la suite des itérés sous la forme $x^{(k+1)} = Bx^{(k)} + c$, où $B \in \mathcal{M}_N(\mathbb{R})$ et $c \in \mathbb{R}^N$ seront choisis de manière à ce que la méthode itérative ainsi définie soit convergente. On appellera ce type de méthode *Méthode I*, et on verra par la suite un choix plus restrictif qu'on appellera *Méthode II*.

Définition 1.24 (Méthode I) *On appelle méthode itérative de type I pour la résolution du système linéaire* (1.1.1) *une méthode itérative où la suite des itérés $(x^{(k)})_{k\in\mathbb{N}}$ est donnée par :*

$$\begin{cases} \textit{Initialisation} & x^{(0)} \in \mathbb{R}^N \\ \textit{Itération } n & x^{(k+1)} = Bx^{(k)} + c. \end{cases}$$

où $B \in \mathcal{M}_N(\mathbb{R})$ et $c \in \mathbb{R}^N$.

Remarque 1.25 (Condition nécessaire de convergence) *Une condition nécessaire pour que la méthode I converge est que $c = (Id - B)A^{-1}b$. En effet, supposons que la méthode converge. En passant à la limite lorsque n tend vers l'infini sur l'itération n de l'algorithme, on obtient $x = Bx + c$ et comme $x = A^{-1}b$, ceci entraîne $c = (Id - B)A^{-1}b$.*

Remarque 1.26 (Intérêt pratique) *La "méthode I" est assez peu intéressante en pratique, car il faut calculer $A^{-1}b$, sauf si $(Id - B)A^{-1} = \alpha Id$, avec $\alpha \in \mathbb{R}$. On obtient dans ce cas :*

$$\begin{aligned} & B = -\alpha A + Id \\ \textit{et} \quad & c = \alpha b \end{aligned}$$

c'est–à–dire

$$x^{n+1} = x^n + \alpha(b - Ax^n).$$

Le terme $b - Ax^n$ est appelé résidu et la méthode s'appelle dans ce cas la méthode d'extrapolation de Richardson.

Théorème 1.27 (Convergence de la méthode de type I) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ A inversible, $b \in \mathbb{R}^N$. On considère la méthode I avec $B \in \mathcal{M}_N(\mathbb{R})$ et*

$$c = (Id - B)A^{-1}b. \tag{1.5.28}$$

Alors la méthode I converge si et seulement si le rayon spectral $\rho(B)$ de la matrice B vérifie $\rho(B) < 1$.

Démonstration
Soit $B \in \mathcal{M}_N(\mathbb{R})$.
Soit x la solution du système linéaire (1.1.1) ; grâce à (1.5.28), $x = Bx + c$, et comme $x^{(k+1)} = Bx^{(k)} + c$, on a donc $x^{(k+1)} - x = B(x^{(k)} - x)$ et par récurrence sur k,

$$x^{(k)} - x = B^k(x^{(0)} - x), \quad \forall k \in \mathbb{N}. \tag{1.5.29}$$

On en déduit par le lemme 1.5 que la méthode converge si et seulement si $\rho(B) < 1$. En effet

($\Rightarrow$) Supposons que $\rho(B) \geq 1$ et montrons que la méthode I ne converge pas. Si $\rho(B) \geq 1$ il existe $y \in \mathbb{R}^N$ tel que $B^k y \underset{k\to+\infty}{\not\longrightarrow} 0$.

En choisissant $x^{(0)} = x + y = A^{-1}b + y$, l'égalité (1.5.29) devient : $x^{(k)} - x = B^k y \underset{k\to+\infty}{\not\longrightarrow} 0$ par hypothèse et donc la méthode n'est pas convergente.

($\Leftarrow$) Supposons maintenant que $\rho(B) < 1$ alors l'égalité (1.5.29) donne

$$x^{(k)} - x = B^k(x^{(0)} - x) \underset{k\to+\infty}{\to} 0$$

car $\rho(B) < 1$. Donc $x^{(k)} \underset{k\to+\infty}{\longrightarrow} x = A^{-1}b$. La méthode est bien convergente.

■

Définition 1.28 (Méthode II) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, $b \in \mathbb{R}^N$. Soient $\tilde{M}$ et $\tilde{N} \in \mathcal{M}_N(\mathbb{R})$ des matrices telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible (et facile à inverser).*
On appelle méthode de type II pour la résolution du système linéaire (1.1.1) *une méthode itérative où la suite des itérés $(x^{(k)})_{k\in\mathbb{N}}$ est donnée par :*

$$\begin{cases} \textit{Initialisation} & x^{(0)} \in \mathbb{R}^N \\ \textit{Itération } n & \tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b. \end{cases} \tag{1.5.30}$$

Remarque 1.29 *Si $\tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b$ pour tout $k \in \mathbb{N}$ et $x^{(k)} \longrightarrow y$ quand $n \longrightarrow +\infty$ alors $\tilde{M}y = \tilde{N}y + b$, c.à.d. $(\tilde{M} - \tilde{N})y = b$ et donc $Ay = b$. En conclusion, si la méthode de type II converge, alors elle converge bien vers la solution du système linéaire.*

Théorème 1.30 (Convergence de la méthode II) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, $b \in \mathbb{R}^N$. Soient $\tilde{M}$ et $\tilde{N} \in \mathcal{M}_N(\mathbb{R})$ des matrices telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible. Alors :*

1. *La méthode définie par* (1.5.30) *converge si et seulement si $\rho(\tilde{M}^{-1}\tilde{N}) < 1$.*
2. *La méthode itérative définie par* (1.5.30) *converge si et seulement si il existe une norme induite notée $\|\cdot\|$ telle que $\|\tilde{M}^{-1}\tilde{N}\| < 1$.*

Démonstration

1. Pour démontrer le point 1, il suffit de réécrire la méthode II avec le formalisme de la méthode I. En effet,
$$\tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b \Longleftrightarrow x^{(k+1)} = \tilde{M}^{-1}\tilde{N}x^{(k)} + \tilde{M}^{-1}b = Bx^{(k)} + c,$$
avec $B = \tilde{M}^{-1}\tilde{N}$ et $c = \tilde{M}^{-1}b$. icicicicicici
2. Si il existe une norme induite notée $\|\cdot\|$ telle que $\|\tilde{M}^{-1}\tilde{N}\| < 1$, alors en vertu du lemme 1.5, $\rho(\tilde{M}^{-1}\tilde{N}) < 1$ et donc la méthode converge ce qui précède.
Réciproquement, si la méthode converge alors $\rho(\tilde{M}^{-1}\tilde{N}) < 1$, et donc il existe $\eta > 0$ tel que $\rho(\tilde{M}^{-1}\tilde{N}) = 1 - \eta$. varepsilon Prenons maintenant $\varepsilon = \frac{\eta}{2}$ et appliquons la proposition 1.7 : il existe une norme induite $\|\cdot\|$ telle que $\|\tilde{M}^{-1}\tilde{N}\| \leq \rho(\tilde{M}^{-1}\tilde{N}) + \varepsilon < 1$, ce qui démontre le résultat.

■

Pour trouver des méthodes itératives de résolution du système (1.1.1), on cherche donc une décomposition de la matrice A de la forme : $A = \tilde{M} - \tilde{N}$, où $\tilde{M}$ est inversible, telle que le système $\tilde{M}y = d$ soit un système facile à résoudre (par exemple $\tilde{M}$ soit triangulaire).

Théorème 1.31 (Condition suffisante de convergence, méthode II) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive, et soient $\tilde{M}$ et $\tilde{N} \in \mathcal{M}_N(\mathbb{R})$ telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible. Si la matrice $\tilde{M}^t + \tilde{N}$ est symétrique définie positive alors $\rho(\tilde{M}^{-1}\tilde{N}) < 1$, et donc la méthode II converge.*

Démonstration On rappelle (voir exercice (5) page 37) que si $B \in \mathcal{M}_N(\mathbb{R})$, et si $\|\cdot\|$ est une norme induite sur $\mathcal{M}_N(\mathbb{R})$ par une norme sur $\mathbb{R}^N$, on a toujours $\rho(B) \leq \|B\|$. On va donc chercher une norme sur $\mathbb{R}^N$, notée $\|\cdot\|_*$ telle que

$$\|\tilde{M}^{-1}\tilde{N}\|_* = \max\{\|\tilde{M}^{-1}\tilde{N}x\|_*,\ x \in \mathbb{R}^N,\ \|x\|_* = 1\} < 1,$$

(où on désigne encore par $\|.\|_*$ la norme induite sur $\mathcal{M}_N(\mathbb{R})$) ou encore :

$$\|\tilde{M}^{-1}\tilde{N}x\|_* < \|x\|_*,\ \forall x \in \mathbb{R}^N,\ x \neq 0. \tag{1.5.31}$$

On définit la norme $\|\cdot\|_*$ par $\|x\|_* = \sqrt{Ax \cdot x}$, pour tout $x \in \mathbb{R}^N$. Comme A est symétrique définie positive, $\|\cdot\|_*$ est bien une norme sur $\mathbb{R}^N$, induite par le produit scalaire $(x|y)_A = Ax \cdot y$. On va montrer que la propriété (1.5.31) est vérifiée par cette norme. Soit $x \in \mathbb{R}^N$, $x \neq 0$, on a : $\|\tilde{M}^{-1}\tilde{N}x\|_*^2 = A\tilde{M}^{-1}\tilde{N}x \cdot \tilde{M}^{-1}\tilde{N}x$. Or $\tilde{N} = \tilde{M} - A$, et donc : $\|\tilde{M}^{-1}\tilde{N}x\|_*^2 = A(Id - \tilde{M}^{-1}A)x \cdot (Id - \tilde{M}^{-1}A)x$. Soit $y = \tilde{M}^{-1}Ax$; remarquons que $y \neq 0$ car $x \neq 0$ et $\tilde{M}^{-1}A$ est inversible. Exprimons $\|\tilde{M}^{-1}\tilde{N}x\|_*^2$ à l'aide de y.

$$\|\tilde{M}^{-1}\tilde{N}x\|_*^2 = A(x-y)\cdot(x-y) = Ax\cdot x - 2Ax\cdot y + Ay\cdot y = \|x\|_*^2 - 2Ax\cdot y + Ay\cdot y.$$

Pour que $\|\tilde{M}^{-1}\tilde{N}x\|_*^2 < \|x\|_*^2$ (et par suite $\rho(\tilde{M}^{-1}\tilde{N}) < 1$), il suffit donc de montrer que $-2Ax\cdot y + Ay\cdot y < 0$. Or, comme $\tilde{M}y = Ax$, on a : $-2Ax\cdot y + Ay\cdot y = -2\tilde{M}y\cdot y + Ay\cdot y$. En écrivant : $\tilde{M}y\cdot y = y\cdot\tilde{M}^t y = \tilde{M}^t y\cdot y$, on obtient donc que : $-2Ax\cdot y + Ay\cdot y = (-\tilde{M} - \tilde{M}^t + A)y\cdot y$, et comme $A = \tilde{M} - \tilde{N}$ on obtient $-2Ax\cdot y + Ay\cdot y = -(\tilde{M}^t + \tilde{N})y\cdot y$. Comme $\tilde{M}^t + \tilde{N}$ est symétrique définie positive par hypothèse et que $y \neq 0$, on en déduit que $-2Ax\cdot y + Ay\cdot y < 0$, ce qui termine la démonstration. ■

Estimation de la vitesse de convergence On montre dans l'exercice 43 page 52 que si la suite $(x^{(k)})_{k\in\mathbb{N}} \subset \mathbb{R}$ est donnée par une "méthode I" (voir définition 1.24 page 28) convergente, *i.e.* $x^{(k+1)} = Bx^{(k)} + C$ (avec $\rho(B) < 1$), et si on suppose que x est la solution du système (1.1.1), et que $x^{(k)} \to x$ quand $k \to \infty$, alors $\dfrac{\|x^{(k+1)} - x\|}{\|x^{(k)} - x\|} \longrightarrow \rho(B)$ quand $k \to +\infty$ (sauf cas particuliers) indépendamment de la norme sur $\mathbb{R}^N$. Le rayon spectral $\rho(B)$ de la matrice B est donc une bonne estimation de la vitesse de convergence. Pour estimer cette vitesse de convergence lorsqu'on ne connaît pas x, on peut utiliser le fait (voir encore l'exercice 43 page 52) qu'on a aussi

$$\frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)} - x^{(k-1)}\|} \longrightarrow \rho(B) \quad \text{lorsque } n \to +\infty,$$

ce qui permet d'évaluer la vitesse de convergence de la méthode par le calcul des itérés courants.

1.5.3 Méthodes de Jacobi, Gauss-Seidel et SOR/SSOR

Décomposition par blocs de A :

Dans de nombreux cas pratiques, les matrices des systèmes linéaires à résoudre ont une structure "par blocs", et on se sert de cette structure lors de la résolution par une méthode itérative.

Définition 1.32 *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. Une décomposition par blocs de A est définie par un entier $S \leq N$, des entiers $(n_i)_{i=1,\dots,S}$ tels que $\sum_{i=1}^S n_i = N$, et S^2 matrices $A_{i,j} \in \mathcal{M}_{n_i,n_j}(\mathbb{R})$ (ensemble des matrices rectangulaires à n_i lignes et n_j colonnes, telles que les matrices $A_{i,i}$ soient inversibles pour $i = 1,\dots,S$ et*

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & & \dots & A_{1,S} \\ A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & A_{S-1,S} \\ A_{S,1} & \dots & & \dots & A_{S,S-1} & A_{S,S} \end{bmatrix} \tag{1.5.32}$$

Remarque 1.33

1. *Si $S = N$ et $n_i = 1\ \forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient.*
2. *Si A est symétrique définie positive, la condition $A_{i,i}$ inversible dans la définition 1.32 est inutile car $A_{i,i}$ est nécessairement symétrique définie positive donc inversible. Prenons par exemple $i = 1$; soit $y \in \mathbb{R}^{n_1}$, $y \neq 0$ et $x = (y, 0 \dots, 0)^t \in \mathbb{R}^N$. Alors $A_{1,1}y \cdot y = Ax \cdot x > 0$ donc $A_{1,1}$ est symétrique définie positive.*
3. *Si A est une matrice triangulaire par blocs, c.à.d. de la forme* (1.5.32) *avec $A_{i,j} = 0$ si $j > i$, alors*

$$\det(A) = \prod_{i=1}^{S} \det(A_{i,i}).$$

Par contre si A est décomposée en 2×2 blocs carrés (i.e. tels que $n_i = m_j$, $\forall (i,j) \in \{1,2\}$), on a en général : $\det(A) \neq \det(A_{1,1})\det(A_{2,2}) - \det(A_{1,2})\det(A_{2,1})$.

Méthode de Jacobi

On peut remarquer que le choix le plus simple pour le système $\tilde{M}x = d$ soit facile à résoudre (on rappelle que c'est un objectif de la mise sous forme méthode de type II) est de prendre pour $\tilde{M}$ une matrice diagonale. La méthode de Jacobi [4] consiste à prendre pour $\tilde{M}$ la matrice diagonale D formée par les blocs diagonaux de A :

$$D = \begin{bmatrix} A_{1,1} & 0 & \dots & & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \dots & & \dots & 0 & A_{S,S} \end{bmatrix}.$$

Dans la matrice ci-dessus, 0 désigne un bloc nul.

4. Carl Gustav Jakob Jacobi, (Potsdam, 1804 - Berlin, 1851), mathématicien allemand. Issu d'une famille juive, il étudie à l'Université de Berlin, où il obtient son doctorat en 1825, à peine âgé de 21 ans. Sa thèse est une discussion analytique de la théorie des fractions. En 1829, il devient professeur de mathématique à l'Université de Kï¿½nigsberg, et ce jusqu'en 1842. Il fait une dépression, et voyage en Italie en 1843. à son retour, il déménage à Berlin où il sera pensionnaire royal jusqu'à sa mort. Dans une lettre du 2 juillet 1830 adressée à Legendre, Jacobi écrit : ï¿½ M. Fourier avait l'opinion que le but principal des mathématiques était l'utilité publique et l'explication des phénomènes naturels ; mais un philosophe comme lui aurait dû savoir que le but unique de la science, c'est l'honneur de l'esprit humain, et que sous ce titre, une question de nombres vaut autant qu'une question du système du monde. ï¿½ L'expression est restée, et renvoie à un débat toujours d'actualité.

On a alors $\tilde{N} = E + F$, où E et F sont constitués des blocs triangulaires inférieurs et supérieurs de la matrice A :

$$E = \begin{bmatrix} 0 & 0 & \dots & & \dots & 0 \\ -A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ -A_{S,1} & \dots & & \dots & -A_{S,S-1} & 0 \end{bmatrix}$$

et

$$F = \begin{bmatrix} 0 & -A_{1,2} & \dots & & \dots & -A_{1,S} \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & -A_{S-1,S} \\ 0 & \dots & & \dots & 0 & 0 \end{bmatrix}.$$

On a bien $A = \tilde{M} - \tilde{N}$ et avec D, E et F définies comme ci-dessus, la méthode de Jacobi s'écrit :

$$\begin{cases} x^{(0)} \in \mathbb{R}^N \\ Dx^{(k+1)} = (E+F)x^{(k)} + b. \end{cases} \tag{1.5.33}$$

Lorsqu'on écrit la méthode de Jacobi comme une méthode I, on a $B = D^{-1}(E+F)$; on notera J cette matrice. En introduisant la décomposition par blocs de x, solution recherchée de (1.1.1), c.à.d. : $x = [x_1, \dots, x_S]^t$, où $x_i \in \mathbb{R}^{n_i}$, on peut aussi écrire la méthode de Jacobi sous la forme :

$$\begin{cases} x_0 \in \mathbb{R}^N \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, S. \end{cases} \tag{1.5.34}$$

Si $S = N$ et $n_i = 1 \; \forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient, et on obtient la méthode de Jacobi par points (aussi appelée méthode de Jacobi), qui s'écrit donc :

$$\begin{cases} x_0 \in \mathbb{R}^N \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, N. \end{cases} \tag{1.5.35}$$

Méthode de Gauss-Seidel

L'idée de la méthode de Gauss-Seidel[5] est d'utiliser le calcul des composantes de l'itéré $(k+1)$ dès qu'il est effectué. Par exemple, pour calculer la deuxième composante $x_2^{(k+1)}$ du vecteur $x^{(k+1)}$, on pourrait employer la "nouvelle" valeur $x_1^{(k+1)}$ qu'on vient de calculer plutôt que la valeur $x_1^{(k)}$ comme dans (1.5.34) ; de même, dans le calcul de $x_3^{(k+1)}$, on pourrait employer les "nouvelles" valeurs $x_1^{(k+1)}$ et $x_2^{(k+1)}$ plutôt que les valeurs $x_1^{(k)}$ et $x_2^{(k)}$. Cette idée nous suggère de remplacer dans (1.5.34) $x_j^{(k)}$ par $x_j^{(k+1)}$ si $j < i$. On obtient donc l'algorithme suivant :

5. Philipp Ludwig von Seidel (Zweibrï¿½cken, Allemagne 1821 ? Munich, 13 August 1896) mathématicien allemand dont il est dit qu'il a découvert en 1847 le concept crucial de la convergence uniforme en étudiant une démonstration incorrecte de Cauchy.

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i, \quad i = 1, \ldots, S. \end{array} \right. \tag{1.5.36}$$

Notons que l'algorithme de Gauss–Seidel par points (cas ou $S = N$ et $n_i = 1$) s'écrit donc :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k+1)} - \sum_{i<j} a_{i,j}x_j^{(k)} + b_i, \quad i = 1, \ldots, N. \end{array} \right. \tag{1.5.37}$$

La méthode de Gauss–Seidel s'écrit donc sous forme de méthode II avec $\tilde{M} = D - E$ et $\tilde{N} = F$:

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^N \\ (D - E)x^{(k+1)} = Fx^{(k)} + b. \end{array} \right. \tag{1.5.38}$$

Lorsqu'on écrit la méthode de Gauss–Seidel comme une méthode I, on a $B = (D - E)^{-1}F$; on notera $\mathcal{L}_1$ cette matrice, dite matrice de Gauss-Seidel.

Méthodes SOR et SSOR

L'idée de la méthode de sur-relaxation (SOR = Successive Over Relaxation) est d'utiliser la méthode de Gauss-Seidel pour calculer un itéré intermédiaire $\tilde{x}^{(k+1)}$ qu'on "relaxe" ensuite pour améliorer la vitesse de convergence de la méthode. On se donne $0 < \omega < 2$, et on modifie l'algorithme de Gauss–Seidel de la manière suivante :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^N \\ A_{i,i}\tilde{x}_i^{(k+1)} = -\displaystyle\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1 - \omega)x_i^{(k)}, \quad i = 1, \ldots, S. \end{array} \right. \tag{1.5.39}$$

(Pour $\omega = 1$ on retrouve la méthode de Gauss–Seidel.)
L'algorithme ci-dessus peut aussi s'écrire (en multipliant par $A_{i,i}$ la ligne 3 de l'algorithme (1.5.39)) :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ A_{i,i}x_i^{(k+1)} = \omega\left[-\displaystyle\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i\right] \\ \qquad\qquad +(1 - \omega)A_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.5.40}$$

On obtient donc

$$(D - \omega E)x^{(k+1)} = \omega Fx^{(k)} + \omega b + (1 - \omega)Dx^{(k)}.$$

L'algorithme SOR s'écrit donc comme une méthode II avec

$$\tilde{M} = \frac{D}{\omega} - E \text{ et } \tilde{N} = F + \left(\frac{1 - \omega}{\omega}\right) D.$$

Il est facile de vérifier que $A = \tilde{M} - \tilde{N}$.

L'algorithme SOR s'écrit aussi comme une méthode I avec

$$B = \left(\frac{D}{\omega} - E\right)^{-1} (F + \left(\frac{1 - \omega}{\omega}\right) D).$$

On notera $\mathcal{L}_\omega$ cette matrice.

Remarque 1.34 (Méthode de Jacobi relaxée) *On peut aussi appliquer une procédure de relaxation avec comme méthode iérative "de base" la méthode de Jacobi, voir à ce sujet l'exercice 36 page 48). Cette méthode est toutefois beaucoup moins employée en pratique (car moins efficace) que la méthode SOR.*

En "symétrisant" le procédé de la méthode SOR, c.à.d. en effectuant les calculs SOR sur les blocs dans l'ordre 1 à N puis dans l'ordre N à 1, on obtient la méthode de sur-relaxation symétrisée (SSOR = Symmetric Successive Over Relaxation) qui s'écrit dans le formalisme de la méthode I avec

$$B = \underbrace{\left(\frac{D}{\omega} - F\right)^{-1}\left(E + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \ S\ldots 1}\underbrace{\left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \ 1\ldots S}.$$

Etude théorique de convergence

On aimerait pouvoir répondre aux questions suivantes :

1. Les méthodes sont–elles convergentes ?
2. Peut-on estimer leur vitesse de convergence ?
3. Peut-on estimer le coefficient de relaxation ω optimal dans la méthode SOR, c.à.d. celui qui donnera la plus grande vitesse de convergence ?

On va maintenant donner des réponses, partielles dans certains cas, faute de mieux, à ces questions.

Convergence On rappelle qu'une méthode itérative de type I, *i.e.* écrite sous la forme $x^{(n+1)} = Bx^{(n)} + C$ converge si et seulement si $\rho(B) < 1$ (voir le théorème 1.27 page 28).

Théorème 1.35 (Sur la convergence de la méthode SOR)
Soit $A \in \mathcal{M}_N(\mathbb{R})$ qui admet une décomposition par blocs définie dans la définition 1.5.32 page 31 ; soient D la matrice constituée par les blocs diagonaux, $-E$ (resp. $-F$) la matrice constituée par les blocs triangulaires inférieurs (resp. supérieurs) ; on a donc : $A = D - E - F$. Soit $\mathcal{L}_\omega$ la matrice d'itération de la méthode SOR (et de la méthode de Gauss–Seidel pour $\omega = 1$) définie par :

$$\mathcal{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right), \quad \omega \neq 0.$$

Alors :

1. *Si $\rho(\mathcal{L}_\omega) < 1$ alors $0 < \omega < 2$.*
2. *Si on suppose de plus que A symétrique définie positive, alors :*

$$\rho(\mathcal{L}_\omega) < 1 \text{ si et seulement si } 0 < \omega < 2.$$

En particulier, si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge.

Démonstration du théorème 1.35 :

1. Calculons $\det(\mathcal{L}_\omega)$. Par définition,

$$\mathcal{L}_\omega = \tilde{M}^{-1}\tilde{N}, \text{ avec } \tilde{M} = \frac{1}{\omega}D - E \text{ et } \tilde{N} = F + \frac{1-\omega}{\omega}D.$$

Donc $\det(\mathcal{L}_\omega) = (\det(\tilde{M}))^{-1}\det(\tilde{N})$. Comme $\tilde{M}$ et $\tilde{N}$ sont des matrices triangulaires par blocs, leurs déterminants sont les produits des déterminants des blocs diagonaux (voir la remarque 1.33 page 31). On a donc :

$$\det(\mathcal{L}_\omega) = \frac{(\frac{1-\omega}{\omega})^N \det(D)}{(\frac{1}{\omega})^N \det(D)} = (1-\omega)^N.$$

Or le déterminant d'une matrice est aussi le produit des valeurs propres de cette matrice (comptées avec leur multiplicités algébriques), dont les valeurs absolues sont toutes, par définition, inférieures au rayon spectral. On a donc : $|\det(\mathcal{L}_\omega)| = |(1-\omega)^N| \leq (\rho(\mathcal{L}_\omega))^N$, d'où le résultat.

2. Supposons maintenant que A est une matrice symétrique définie positive, et que $0 < \omega < 2$. Montrons que $\rho(\mathcal{L}_\omega) < 1$. Par le théorème 1.31 page 29, il suffit pour cela de montrer que $\tilde{M}^t + \tilde{N}$ est une matrice symétrique définie positive. Or,

$$\tilde{M}^t = \left(\frac{D}{\omega} - E\right)^t = \frac{D}{\omega} - F,$$

$$\tilde{M}^t + \tilde{N} = \frac{D}{\omega} - F + F + \frac{1-\omega}{\omega}D = \frac{2-\omega}{\omega}D.$$

La matrice $\tilde{M}^t + \tilde{N}$ est donc bien symétrique définie positive.

■

Remarque 1.36 (Comparaison Gauss–Seidel/Jacobi) *On a vu (théorème 1.35) que si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge. Par contre, même dans le cas où A est symétrique définie positive, il existe des cas où la méthode de Jacobi ne converge pas, voir à ce sujet l'exercice 30 page 46.*

Remarquons que le résultat de convergence des méthodes itératives donné par le théorème précédent n'est que partiel, puisqu'il ne concerne que les matrices symétriques définies positives et que les méthodes Gauss-Seidel et SOR. On a aussi un résultat de convergence de la méthode de Jacobi pour les matrices à diagonale dominante stricte, voir exercice 32 page 47, et un résultat de comparaison des méthodes pour les matrices tridiagonales par blocs, voir le théorème 1.37 donné ci-après. Dans la pratique, il faudra souvent compter sur sa bonne étoile...

Estimation du coefficient de relaxation optimal de SOR La question est ici d'estimer le coefficient de relaxation ω optimal dans la méthode SOR, c.à.d. le coefficient $\omega_0 \in]0,2[$ (condition nécessaire pour que la méthode SOR converge, voir théorème 1.35) tel que $\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_\omega)\ \forall \omega \in]0,2[$.
D'après le paragraphe précédent ce ω_0 donnera la meilleure convergence possible pour SOR. On sait le faire dans le cas assez restrictif des matrices tridiagonales par blocs. On ne fait ici qu'énoncer le résultat dont la démonstration est donnée dans le livre de Ph. Ciarlet conseillé en début de cours.

Théorème 1.37 (Coefficient optimal, matrice tridiagonale) *On considère une matrice $A \in \mathcal{M}_N(\mathbb{R})$ qui admet une décomposition par blocs définie dans la définition 1.5.32 page 31 ; on suppose que la matrice A est tridiagonale par blocs, c.à.d. $A_{i,j} = 0$ si $|i-j| > 1$; soient $\mathcal{L}_1$ et J les matrices d'itération respectives des méthodes de Gauss-Seidel et Jacobi, alors :*

1. *$\rho(\mathcal{L}_1) = (\rho(J))^2$: la méthode de Gauss–Seidel converge (ou diverge) donc plus vite que celle de Jacobi.*
2. *On suppose de plus que toutes les valeurs propres de la matrice d'itération J de la méthode de Jacobi sont réelles. alors le paramètre de relaxation optimal, c.à.d. le paramètre ω_0 tel que $\rho(\mathcal{L}_{\omega_0}) = \min\{\rho(\mathcal{L}_\omega), \omega \in]0,2[\}$, s'exprime en fonction du rayon spectral $\rho(J)$ de la matrice J par la formule :*

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(J)^2}} > 1,$$

et on a : $\rho(\mathcal{L}_{\omega_0}) = \omega_0 - 1$.

1.5.4 Recherche de valeurs propres et vecteurs propres

Les techniques de recherche des éléments propres, c.à.d. des valeurs et vecteurs propres (voir Définition 1.4 page 6) d'une matrice sont essentielles dans de nombreux domaines d'application, par exemple en dynamique des structures : la recherche des modes propres d'une structure peut s'avérer importante pour le dimensionnement de structures sous contraintes dynamiques, voir à ce propos l'exemple célèbre du "Tacoma Bridge", décrit dans les livres de M. Braun (en anglais) et M. Schatzman (en français) conseillés en début de cours.
On donne dans les exercices qui suivent deux méthodes assez classiques de recherche de valeurs propres d'une matrice qui sont la méthode de la puissance (exercice 43 page 52) et celui de la puissance inverse (exercice 44 page 53). Citons également une méthode très employée, la méthode QR, qui est présente dans de nombreuses bibliothèques de programmes. On pourra se référer aux ouvrages de Ph. Ciarlet et de M. Schatzman, de D. Serre et de P. Lascaux et R. Theodor (voir introduction).

1.6 Exercices

Exercice 1 (Matrices symétriques définies positives) *Suggestions en page 56, corrigé en page 60.*

On rappelle que toute matrice $A \in \mathcal{M}_N(\mathbb{R})$ symétrique est diagonalisable dans $\mathbb{R}$ (cf. lemme 1.12 page 8). Plus précisément, on a montré en cours que, si $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice symétrique, il existe une base de $\mathbb{R}^N$, notée $\{f_1, \ldots, f_N\}$, et il existe $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ t.q. $Af_i = \lambda_i f_i$, pour tout $i \in \{1, \ldots, N\}$, et $f_i \cdot f_j = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, N\}$ ($x \cdot y$ désigne le produit scalaire de x avec y dans $\mathbb{R}^N$).

1. Soit $A \in \mathcal{M}_N(\mathbb{R})$. On suppose que A est symétrique définie positive, montrer que les éléments diagonaux de A sont strictements positifs.
2. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Montrer que A est symétrique définie positive si et seulement si toutes les valeurs propres de A sont strictement positives.
3. Soit $A \in \mathcal{M}_N(\mathbb{R})$. On suppose que A est symétrique définie positive. Montrer qu'on peut définir une unique matrice $B \in \mathcal{M}_N(\mathbb{R})$, B symétrique définie positive t.q. $B^2 = A$ (on note $B = A^{\frac{1}{2}}$).

Exercice 2 (Normes de l'Identité)

Soit Id la matrice "Identité" de $\mathcal{M}_N(\mathbb{R})$. Montrer que pour toute norme induite on a $\|Id\| = 1$ et que pour toute norme matricielle on a $\|Id\| \geq 1$.

Exercice 3 (Normes induites particulières) *Suggestions en page 56, corrigé détaillé en page 60.*

Soit $A = (a_{i,j})_{i,j \in \{1,\ldots,N\}} \in \mathcal{M}_N(\mathbb{R})$.

1. On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Montrer que $\|A\|_\infty = \max_{i \in \{1,\ldots,N\}} \sum_{j=1}^N |a_{i,j}|$.
2. On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Montrer que $\|A\|_1 = \max_{j \in \{1,\ldots,N\}} \sum_{i=1}^N |a_{i,j}|$.
3. On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_2$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_2$. Montrer que $\|A\|_2 = (\rho(A^t A))^{\frac{1}{2}}$.

Exercice 4 (Norme non induite)

Pour $A = (a_{i,j})_{i,j \in \{1,\ldots,N\}} \in \mathcal{M}_N(\mathbb{R})$, on pose $\|A\|_s = (\sum_{i,j=1}^N a_{i,j}^2)^{\frac{1}{2}}$.

1. Montrer que $\|\cdot\|_s$ est une norme matricielle mais n'est pas une norme induite (pour $N > 1$).

2. Montrer que $\|A\|_s^2 = tr(A^tA)$. En déduire que $\|A\|_2 \leq \|A\|_s \leq \sqrt{N}\|A\|_2$ et que $\|Ax\|_2 \leq \|A\|_s\|x\|_2$, pour tout $A \in \mathcal{M}_N(\mathbb{R})$ et tout $x \in \mathbb{R}^N$.
3. Chercher un exemple de norme non matricielle.

Exercice 5 (Rayon spectral) *Suggestions en page 56, corrigé détaillé en page 61.*

On munit $\mathcal{M}_N(\mathbb{R})$ d'une norme, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_N(\mathbb{R})$.

1. Montrer que $\rho(A) < 1$ si et seulement si $A^k \to 0$ quand $k \to \infty$.

2. Montrer que : $\rho(A) < 1 \Rightarrow \limsup_{k\to\infty} \|A^k\|^{\frac{1}{k}} \leq 1$.

3. Montrer que :$\liminf_{k\to\infty} \|A^k\|^{\frac{1}{k}} < 1 \Rightarrow \rho(A) < 1$.

4. Montrer que $\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}$.

5. On suppose ici que $\|\cdot\|$ est une norme matricielle, déduire de la question précédente que $\rho(A) \leq \|A\|$. (On a ainsi démontré le lemme 1.5).

Exercice 6 (Valeurs propres nulles d'un produit de matrices)

Soient p et n des entiers naturels non nuls tels que $n \leq p$, et soient $A \in \mathcal{M}_{n,p}(\mathbb{R})$ et $B \in \mathcal{M}_{p,n}(\mathbb{R})$. (On rappelle que $\mathcal{M}_{n,p}(\mathbb{R})$ désigne l'ensemble des matrices à n lignes et p colonnes.)

1. Montrer que λ est valeur propre non nulle de AB si et seulement si λ est valeur propre non nulle de BA.
2. Montrer que si $\lambda = 0$ est valeur propre de AB alors λ est valeur propre nulle de BA.
   (Il est conseillé de distinguer les cas $Bx \neq 0$ et $Bx = 0$, où x est un vecteur propre associé à la $\lambda = 0$ valeur propre de AB. Pour le deuxième cas, on pourra distinguer selon que $ImA = \mathbb{R}^n$ ou non.)
3. Montrer en donnant un exemple que λ peut être une valeur propre nulle de BA sans être valeur propre de AB.
   (Prendre par exemple $n = 1$, $p = 2$.)
4. On suppose maintenant que $n = p$, déduire des questions 1. et 2. que l'ensemble des valeurs propres de AB est égal à l'ensemble des valeurs propres de la matrice BA.

Exercice 7 (Rayon spectral) *Corrigé en page 62.*

Soit $A \in \mathcal{M}_N(\mathbb{R})$. Montrer que si A est diagonalisable, il existe une norme induite sur $\mathcal{M}_N(\mathbb{R})$ telle que $\rho(A) = \|A\|$. Montrer par un contre exemple que ceci peut être faux si A n'est pas diagonalisable.

Exercice 8 (Sur le rayon spectral)

On définit les matrices carrées d'ordre 2 suivantes :

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},\ B = \begin{pmatrix} -1 & 0 \\ -1 & -1 \end{pmatrix},\ C = A + B.$$

Calculer le rayon spectral de chacune des matrices A, B et C et en déduire que le rayon spectral ne peut être ni une norme, ni même une semi-norme sur l'espace vectoriel des matrices.

Exercice 9 (Série de Neumann) *Suggestions en page 56, corrigé détaillé en page 62.*

Soient $A \in \mathcal{M}_N(\mathbb{R})$ et $\|\cdot\|$ une norme matricielle.

1. Montrer que si $\rho(A) < 1$, les matrices $Id - A$ et $Id + A$ sont inversibles.

2. Montrer que la série de terme général A^k converge (vers $(Id - A)^{-1}$) si et seulement si $\rho(A) < 1$.

Exercice 10 (Normes matricielles)

Soit $\|.\|$ une norme matricielle quelconque, et soit $A \in \mathcal{M}_N(\mathbb{R})$ telle que $\rho(A) < 1$ (on rappelle qu'on note $\rho(A)$ le rayon spectral de la matrice A). Pour $x \in \mathbb{R}^N$, on définit $\|x\|_*$ par :

$$\|x\|_* = \sum_{j=0}^{\infty} \|A^j x\|.$$

1. Montrer que l'application définie de $\mathbb{R}^N$ dans $\mathbb{R}$ par $x \mapsto \|x\|_*$ est une norme.
2. Soit $x \in \mathbb{R}^N$ tel que $\|x\|_* = 1$. Calculer $\|Ax\|_*$ en fonction de $\|x\|$, et en déduire que $\|A\|_* < 1$.
3. On ne suppose plus que $\rho(A) < 1$. Soit $\varepsilon > 0$ donné. Construire à partir de la norme $\|.\|$ une norme induite $\|.\|_{**}$ telle que $\|A\|_{**} \leq \rho(A) + \varepsilon$.

Exercice 11 (Décomposition LDL^t et LL^t) *Corrigé en page 63*

1. Soit $A = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$.
Calculer la décomposition LDL^t de A. Existe-t-il une décomposition LL^t de A ?

2. Montrer que toute matrice de $\mathcal{M}_N(\mathbb{R})$ symétrique définie positive admet une décomposition LDL^t.

3. Ecrire l'algorithme de décomposition LDL^t. La matrice $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ admet-elle une décomposition LDL^t ?

Exercice 12 (Décomposition LU et LL^t de matrices 3×3)

1. Soit $A = \begin{pmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{pmatrix}$. Donner la décomposition LU de A.

2. Soit $B = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$. Donner la décomposition LL^t de B.

3. Que deviennent les coefficients nuls dans la décomposition LL^t ci-dessus ? Quelle est la propriété vue en cours qui est ainsi vérifiée ?

Exercice 13

Soient a, b, c et d des nombres réels. On considère la matrice suivante :

$$A = \begin{bmatrix} a & a & a & a \\ a & b & b & b \\ a & b & c & c \\ a & b & c & d \end{bmatrix}.$$

Appliquer l'algorithme d'élimination de Gauss à A pour obtenir sa décomposition LU.
Donner les conditions sur a, b, c et d pour que la matrice A soit inversible.

Exercice 14 (Sur la méthode LL^t) *Corrigé détaillé en page 65.*

Soit A une matrice carrée d'ordre N, symétrique définie positive et pleine. On cherche à résoudre le système $A^2x = b$.
On propose deux méthodes de résolution de ce système :

1. Calculer A^2, effectuer la décomposition LL^t de A^2, résoudre le système $LL^tx = b$.
2. Calculer la décomposition LL^t de A, résoudre les systèmes $LL^ty = b$ et $LL^tx = y$.

Calculer le nombre d'opérations élémentaires nécessaires pour chacune des deux méthodes et comparer.

Exercice 15 (Décomposition LL^t d'une matrice tridiagonale symétrique)

Soit $A \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive et tridiagonale (i.e. $a_{i,j} = 0$ si $i - j > 1$).

1. Montrer que A admet une décomposition LL^t, où L est de la forme

$$L = \begin{pmatrix} \alpha_1 & 0 & \dots & & 0 \\ \beta_2 & \alpha_2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \dots & 0 \\ \vdots & \ddots & \ddots & \dots & \vdots \\ 0 & \cdots & 0 & \beta_N & \alpha_N \end{pmatrix}.$$

2. Donner un algorithme de calcul des coefficients α_i et β_i, en fonction des coefficients $a_{i,j}$, et calculer le nombre d'opérations élementaires nécessaires dans ce cas.
3. En déduire la décomposition LL^t de la matrice :

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}.$$

4. L'inverse d'une matrice inversible tridiagonale est elle tridiagonale ?

Exercice 16 (Choleski pour matrice bande) *Suggestions en page 56, corrigé en page 65*

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive.

1. On suppose ici que A est tridiagonale. Estimer le nombre d'opérations de la factorisation LL^t dans ce cas.
2. Même question si A est une matrice bande (c'est-à-dire p diagonales non nulles).
3. En déduire une estimation du nombre d'opérations nécessaires pour la discrétisation de l'équation $-u'' = f$ vue page 23. Même question pour la discrétisation de l'équation $-\Delta u = f$ présentée page 25.

Exercice 17 (Un système par blocs)

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée d'ordre N inversible, $b, c, f \in \mathbb{R}^N$. Soient α et $\gamma \in \mathbb{R}$. On cherche à résoudre le système suivant (avec $x \in \mathbb{R}^N$, $\lambda \in \mathbb{R}$) :

$$\begin{aligned} &Ax + b\lambda = f, \\ &c \cdot x + \alpha\lambda = \gamma. \end{aligned} \tag{1.6.41}$$

1. Ecrire le système (1.6.41) sous la forme : $My = g$, où M est une matrice carrée d'ordre $N+1$, $y \in \mathbb{R}^{N+1}$, $g \in \mathbb{R}^{N+1}$. Donner l'expression de M, y et g.

2. Donner une relation entre A, b, c et α, qui soit une condition nécessaire et suffisante pour que le système (1.6.41) soit inversible. Dans toute la suite, on supposera que cette relation est vérifiée.
3. On propose la méthode suivante pour la résolution du système (1.6.41) :
   (a) Soient z solution de $Az = b$, et h solution de $Ah = f$.
   (b) $x = h - \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z} z, \qquad \lambda = \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z}.$

   Montrer que $x \in \mathbb{R}^N$ et $\lambda \in \mathbb{R}$ ainsi calculés sont bien solutions du système (1.6.41).
4. On suppose dans cette question que A est une matrice bande, dont la largeur de bande est p.
   (a) Calculer le coût de la méthode de résolution proposée ci-dessus en utilisant la méthode LU pour la résolution des systèmes linéaires.
   (b) Calculer le coût de la résolution du système $My = g$ par la méthode LU (en profitant ici encore de la structure creuse de la matrice A).
   (c) Comparer et conclure.
   Dans les deux cas, le terme d'ordre supérieur est $2Nq^2$, et les coûts sont donc comparables.

Exercice 18 (Propriétés générales du conditionnement) *Corrigé détaillé en page 67.*

Partie I
On munit $\mathbb{R}^N$ d'une norme, notée $\|\cdot\|$, et $\mathcal{M}_N(\mathbb{R})$ de la norme induite, notée aussi $\|\cdot\|$. Pour une matrice inversible $A \in \mathcal{M}_N(\mathbb{R})$, on note $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$.

1. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. Montrer que $\mathrm{cond}(A) \geq 1$.
2. Montrer que $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$ pour tout $\alpha \in \mathbb{R}^\star$.
3. Soit $A, B \in \mathcal{M}_N(\mathbb{R})$ deux matrices inversibles. Montrer que $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.

Partie II
On suppose maintenant que $\mathbb{R}^N$ est muni de la norme euclidienne usuelle $\|\cdot\| = \|\cdot\|_2$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite (notée aussi $\|\cdot\|_2$. On note alors $\mathrm{cond}_2(A)$ conditionnement d'une matrice A inversible.

1. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On note σ_N [resp. σ_1] la plus grande [resp. petite] valeur propre de A^tA (noter que A^tA est une matrice symétrique définie positive). Montrer que $\mathrm{cond}_2(A) = \sqrt{\sigma_N/\sigma_1}$.
2. On suppose maintenant que A est symétrique définie positive, montrer que $\mathrm{cond}_2(A) = \lambda_N/\lambda_1$ où λ_N [resp. λ_1] est la plus grande [resp. petite] valeur propre de A.
3. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. Montrer que $\mathrm{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^\star$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).
4. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Montrer que $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.
5. Soit $A, B \in \mathcal{M}_N(\mathbb{R})$ deux matrices symétriques définies positives. Montrer que $\mathrm{cond}_2(A+B) \leq \max\{\mathrm{cond}_2(A), \mathrm{cond}_2(B)$

Exercice 19 (Minoration du conditionnement)

Soit $\|\,.\,\|$ une norme induite sur $\mathcal{M}_N(\mathbb{R})$ et soit $A \in \mathcal{M}_N(\mathbb{R})$ telle que $\det(A) \neq 0$.

1. Montrer que si $\|A - B\| < \frac{1}{\|A^{-1}\|}$, alors B est inversible.
2. Montrer que $\mathrm{cond}\,(A) \geq \sup_{\substack{B \in \mathcal{M}_N(\mathbb{R}) \\ detB=0}} \frac{\|A\|}{\|A-B\|}$

Exercice 20 (Minoration du conditionnement) *Corrigé détaillé en page 68.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_N(\mathbb{R})$. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée inversible, $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$ le conditionnement de A, et soit $\delta A \in \mathcal{M}_N(\mathbb{R})$.

1. Montrer que si $A + \delta A$ est singulière, alors
$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta A\|}. \tag{1.6.42}$$
2. On suppose dans cette question que la norme $\|\cdot\|$ est la norme induite par la norme euclidienne sur $\mathbb{R}^N$. Montrer que la minoration (1.6.42) est optimale, c'est-à-dire qu'il existe $\delta A \in \mathcal{M}_N(\mathbb{R})$ telle que $A + \delta A$ soit singulière et telle que l'égalité soit vérifiée dans (1.6.42).
[*On pourra chercher δA de la forme*
$$\delta A = -\frac{y\, x^t}{x^t\, x},$$
avec $y \in \mathbb{R}^N$ convenablement choisi et $x = A^{-1}y$.]
3. On suppose ici que la norme $\|\cdot\|$ est la norme induite par la norme infinie sur $\mathbb{R}^N$. Soit $\alpha \in]0,1[$. Utiliser l'inégalité (1.6.42) pour trouver un minorant, qui tend vers $+\infty$ lorsque α tend vers 0, de $\mathrm{cond}(A)$ pour la matrice
$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & \alpha & -\alpha \\ 1 & \alpha & \alpha \end{pmatrix}.$$

Exercice 21 (Conditionnement du carré)

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice telle que $\det A \neq 0$.

1. Quelle relation existe-t-il en général entre $\mathrm{cond}\,(A^2)$ et $(\mathrm{cond}\,A)^2$?
2. On suppose que A symétrique. Montrer que $\mathrm{cond}_2\,(A^2) = (\mathrm{cond}_2\,A)^2$.
3. On suppose que $\mathrm{cond}_2\,(A^2) = (\mathrm{cond}_2\,A)^2$. Peut-on conclure que A est symétrique ? (justifier la réponse.)

Exercice 22 (Calcul de l'inverse d'une matrice et conditionnement) *Corrigé détaillé en page 69.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_N(\mathbb{R})$. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée inversible. On cherche ici des moyens d'évaluer la précision de calcul de l'inverse de A.

1. On suppose qu'on a calculé B, approximation (en raison par exemple d'erreurs d'arrondi) de la matrice A^{-1}. On pose :
$$\left\{ \begin{array}{lllll} e_1 &=& \dfrac{\|B - A^{-1}\|}{\|A^{-1}\|}, & e_2 &= \dfrac{\|B^{-1} - A\|}{\|A\|} \\ \\ e_3 &=& \|AB - Id\|, & e_4 &= \|BA - Id\| \end{array} \right. \tag{1.6.43}$$
   (a) Expliquer en quoi les quantités e_1, e_2, e_3 et e_4 mesurent la qualité de l'approximation de A^{-1}.
   (b) On suppose ici que $B = A^{-1} + E$, où $\|E\| \leq \varepsilon\|A^{-1}\|$, et que
$$\varepsilon \mathrm{cond}(A) < 1.$$
   Montrer que dans ce cas,
$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon \mathrm{cond}(A)}{1 - \varepsilon \mathrm{cond}(A)},\ e_3 \leq \varepsilon \mathrm{cond}(A) \text{ et } e_4 \leq \varepsilon \mathrm{cond}(A).$$

(c) On suppose maintenant que $AB - Id = E'$ avec $\|E'\| \leq \varepsilon < 1$. Montrer que dans ce cas :

$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon}{1-\varepsilon},\ e_3 \leq \varepsilon \text{ et } e_4 \leq \varepsilon \mathrm{cond}(A).$$

2. On suppose maintenant que la matrice A n'est connue qu'à une certaine matrice d'erreurs près, qu'on note δ_A.

(a) Montrer que la matrice $A + \delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.

(b) Montrer que si la matrice $A + \delta_A$ est inversible,

$$\frac{\|(A+\delta_A)^{-1} - A^{-1}\|}{\|(A+\delta_A)^{-1}\|} \leq \mathrm{cond}(A)\frac{\|\delta_A\|}{\|A\|}.$$

Exercice 23 (Discrétisation)

On considère la discrétisation à pas constant par le schéma aux différences finies symétrique à trois points (vu en cours) du problème (1.4.21) page 23, avec $f \in C([0,1])$. Soit $N \in \mathbb{N}^\star$, N impair. On pose $h = 1/(N+1)$. On note u est la solution exacte, $x_i = ih$, pour $i = 1, \ldots, N$ les points de discrétisation, et $(u_i)_{i=1,\ldots,N}$ la solution du système discrétisé (1.4.23).

1. Montrer que si $u \in C^4([0,1]$, alors la propriété (1.4.22) est vérifiée, c.à.d. :

$$-\frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} = -u''(x_i) + R_i \text{ avec } |R_i| \leq \frac{h^2}{12}\|u^{(4)}\|_\infty.$$

2. Montrer que si f est constante, alors

$$\max_{1\leq i\leq N} |u_i - u(x_i)| = 0.$$

.

3. Soit N fixé, et $\max_{1\leq i\leq N} |u_i - u(x_i)| = 0$. A-t-on forcément que f est constante sur $[0,1]$? (justifier la réponse.)

Exercice 24 (IP-matrice) *Corrigé en page 70*

Soit $N \in \mathbb{N}^\star$, on note $\mathcal{M}_N(\mathbb{R})$ l'ensemble des matrices de N lignes et N colonnes et à coefficients réels. Si $x \in \mathbb{R}^N$, on dit que $x \geq 0$ [resp. $x > 0$] si toutes les composantes de x sont positives [resp. strictement positives]. Soit $A \in \mathcal{M}_N(\mathbb{R})$, on dit que A est une IP-matrice si elle vérifie la propriété suivante :

$$\text{Si } x \in \mathbb{R}^N \text{ est tel que } Ax \geq 0, \text{ alors } x \geq 0,$$

ce qui peut encore s'écrire : $\{x \in \mathbb{R}^N \text{ t.q. } Ax \geq 0\} \subset \{x \in \mathbb{R}^N \text{ t.q. } x \geq 0\}$.

1. Soit $A = (a_{i,j})_{i,j=1,\ldots,N} \in \mathcal{M}_N(\mathbb{R})$. Montrer que A est une IP-matrice si et seulement si A est inversible et $A^{-1} \geq 0$ (c'est-à-dire que tous les coefficients de A^{-1} sont positifs).

2. Soit $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ une matrice réelle d'ordre 2. Montrer que A est une IP-matrice si et seulement si :

$$\left\{\begin{array}{l} ad < bc, \\ a < 0,\ d < 0 \\ b \geq 0,\ c \geq 0 \end{array}\right. \text{ ou } \left\{\begin{array}{l} ad > bc, \\ a > 0,\ d > 0, \\ b \leq 0,\ c \leq 0. \end{array}\right. \tag{1.6.44}$$

3. Montrer que si $A \in \mathcal{M}_N(\mathbb{R})$ est une IP-matrice alors A^t (la transposée de A) est une IP-matrice.

4. Montrer que si A est telle que

$$\begin{aligned} &a_{i,j} \leq 0, \text{ pour tout } i,j = 1,\dots,N,\ i \neq j, \\ &a_{i,i} > \sum_{\substack{j=1 \\ j\neq i}}^{N} |a_{i,j}|, \text{ pour tout } i = 1,\dots,N, \end{aligned} \tag{1.6.45}$$

alors A est une IP-matrice.

5. En déduire que si A est telle que

$$\begin{aligned} &a_{i,j} \leq 0, \text{ pour tout } i,j = 1,\dots,N,\ i \neq j, \\ &a_{i,i} > \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{j,i}|, \text{ pour tout } i = 1,\dots,N, \end{aligned} \tag{1.6.46}$$

alors A est une IP-matrice.

6. Montrer que si $A \in \mathcal{M}_N(\mathbb{R})$ est une IP-matrice et si $x \in \mathbb{R}^N$ alors :

$$Ax > 0 \Rightarrow x > 0.$$

c'est-à-dire que $\{x \in \mathbb{R}^N \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^N \text{ t.q. } x > 0\}$

7. Montrer, en donnant un exemple, qu'une matrice A de $\mathcal{M}_N(\mathbb{R})$ peut vérifier $\{x \in \mathbb{R}^N \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^N \text{ t.q. } x > 0\}$ et ne pas être une IP-matrice.

8. On suppose dans cette question que $A \in \mathcal{M}_N(\mathbb{R})$ est inversible et que $\{x \in \mathbb{R}^N \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^N \text{ t.q. } x > 0\}$. Montrer que A est une IP-matrice.

9. Soit E l'espace des fonctions continues sur $\mathbb{R}$ et admettant la même limite finie en $+\infty$ et $-\infty$. Soit $\mathcal{L}(E)$ l'ensemble des applications linéaires continues de E dans E. Pour $f \in E$, on dit que $f > 0$ (resp. $f \geq 0$) si $f(x) > 0$ (resp. $f(x) \geq 0$) pour tout $x \in \mathbb{R}$. Montrer qu'il existe $T \in \mathcal{L}(E)$ tel que $Tf \geq 0 \Longrightarrow f \geq 0$, et $g \in E$ tel que $Tg > 0$ et $g \not> 0$ (ceci démontre que le raisonnement utilisé en 2 (b) ne marche pas en dimension infinie).

Exercice 25 (M-matrice)

Dans ce qui suit, toutes les inégalités écrites sur des vecteurs ou des matrices sont à entendre au sens composante par composante.
Soit $A = (a_{i,j})_{i,j=1,\dots,n}$ une matrice carrée d'ordre n. On dit que A est une M-matrice si A est une IP-matrice (A est inversible et $A^{-1} \geq 0$, voir exercice 24) qui vérifie de plus

(a) $a_{i,i} > 0$ pour $i = 1,\dots,n$;

(b) $a_{i,j} \leq 0$ pour $i,j = 1,\dots,n, i \neq j$;

1. Soit A une IP-matrice ; montrer que A est une M-matrice si et seulement si la propriété (b) est vérifiée.
2. Soit $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ une matrice réelle d'ordre 2. Montrer que A est une M-matrice si et seulement si :

$$\begin{cases} ad > bc, \\ a > 0,\ d > 0, \\ b \leq 0,\ c \leq 0. \end{cases} \tag{1.6.47}$$

3. Les matrices $A = \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}$ et $B = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ sont-elles des IP-matrices ? des M-matrices ?

4. Soit A la matrice carrée d'ordre 3 définie par :

$$A = \begin{pmatrix} 2 & -1 & \frac{1}{2} \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix}$$

Montrer que $A^{-1} \geq 0$ mais que A n'est pas une M–matrice.

5. Soit A une matrice carrée d'ordre $n = m + p$, avec $m, p \in \mathbb{N}$ tels que $m \geq 1$ et $p \geq 1$, vérifiant :

$$\left.\begin{array}{l} a_{i,i} \geq 0, \\ a_{i,j} \leq 0, \text{ pour } j = 1, \ldots, n, \ j \neq i, \\ a_{i,i} + \displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j} = 0 \end{array}\right\} \text{ pour } i = 1, \ldots, m, \tag{1.6.48}$$

$$\left.\begin{array}{l} a_{i,i} = 1, \\ a_{i,j} = 0, \text{ pour } j = 1, \ldots, n, \ j \neq i, \end{array}\right\} \text{ pour } i = m+1, \ldots, n. \tag{1.6.49}$$

$$\forall i \leq m, \ \exists (k_\ell)_{\ell=1,\ldots,L_i}; k_1 = i, k_{L_i} > m, \text{ et } a_{k_\ell, k_{\ell+1}} < 0, \forall\, \ell = 1, \ldots, L_i. \tag{1.6.50}$$

Soit $b \in \mathbb{R}^n$ tel que $b_i = 0$ pour $i = 1, \ldots, m$. On considère le système linéaire

$$Au = b \tag{1.6.51}$$

5.1 Montrer que le système (1.6.51) admet une et une seule solution.

5.2 Montrer que u est tel que $\min_{k=m+1,n} b_k \leq u_i \leq \max_{k=m+1,n} b_k$. (On pourra pour simplifier supposer que les équations sont numérotées de telle sorte que $\min_{k=m+1,n} b_k = b_{m+2} \leq b_2 \leq \ldots \leq b_n = \max_{k=m+1,n} b_k$.)

6. On considère le problème de Dirichlet suivant :

$$-u'' = 0 \text{ sur } [0,1] \tag{1.6.52a}$$
$$u(0) = -1 \tag{1.6.52b}$$
$$u(1) = 1. \tag{1.6.52c}$$

6.1 Calculer la solution exacte u de ce problème et vérifier qu'elle reste comprise entre -1 et 1.

6.2 Soit $m > 1$ et soient A et b et la matrice et le second membre du système linéaire d'ordre $n = m+2$ obtenu par discrétisation par différences finies avec un pas uniforme $h = \frac{1}{m}$ du problème (1.6.52) (en écrivant les conditions aux limites dans le système). Montrer que la solution $U \in \mathbb{R}^n$ du système $AU = b$ vérifie $-1 \leq u_i \leq 1$.

Exercice 26 (Valeurs propres du Laplacien discret 1D) *Suggestions en page 57, corrigé détaillé en page 71.*

Soit $f \in C([0,1])$. Soit $N \in \mathbb{N}^\star$, N impair. On pose $h = 1/(N+1)$. Soit A la matrice définie par (1.4.24) page 24, issue d'une discrétisation par différences finies (vue en cours) du problème (1.4.21) page 23.
1. Montrer que A est symétrique définie positive.
2. Calculer les valeurs propres et les vecteurs propres de A. [*On pourra commencer par chercher* $\lambda \in \mathbb{R}$ *et* $\varphi \in C^2(\mathbb{R}, \mathbb{R})$ *(φ non identiquement nulle) t.q.* $-\varphi''(x) = \lambda\varphi(x)$ *pour tout* $x \in]0,1[$ *et* $\varphi(0) = \varphi(1) = 0$].
3. Calculer $\mathrm{cond}_2(A)$ et montrer que $h^2 \mathrm{cond}_2(A) \to \frac{4}{\pi^2}$ lorsque $h \to 0$.

Exercice 27 (Conditionnement, réaction diffusion 1d.)

On s'intéresse au conditionnement pour la norme euclidienne de la matrice issue d'une discrétisation par Différences Finies du problème aux limites suivant :

$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.6.53}$$

Soit $N \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,N}$ une "valeur approchée" de la solution u du problème (1.6.53) aux points $\left(\frac{j}{N+1}\right)_{j=1,\ldots,N}$. On rappelle que la discrétisation par différences finies de ce problème consiste à chercher U comme solution du système linéaire $AU = \left(f(\frac{j}{N+1})\right)_{j=1,\ldots,N}$ où la matrice $A \in \mathcal{M}_N(\mathbb{R})$ est définie par $A = (N+1)^2 B + Id$, Id désigne la matrice identité et

$$B = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix}$$

1. (Valeurs propres de la matrice B.)
   On rappelle que le problème aux valeurs propres

   $$\begin{array}{l} -u''(x) = \lambda u(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.6.54}$$

   admet la famille $(\lambda_k, u_k)_{k\in\mathbb{N}^*}$, $\lambda_k = (k\pi)^2$ et $u_k(x) = \sin(k\pi x)$ comme solution. Montrer que les vecteurs $U_k = \left(u_k(\frac{j}{N+1})\right)_{j=1,\ldots,N}$ sont des vecteurs propres de la matrice B. En déduire toutes les valeurs propres de la matrice B.
2. En déduire les valeurs propres de la matrice A.
3. En déduire le conditionnement pour la norme euclidienne de la matrice A.

Exercice 28 (Conditionnement "efficace".) *Suggestions en page 57, corrigé en page 73.*

Soit $f \in C([0,1])$. Soit $N \in \mathbb{N}^\star$, N impair. On pose $h = 1/(N+1)$. Soit A la matrice définie par (1.4.24) page 24, issue d'une discrétisation par différences finies (vue en cours) du problème (1.4.21) page 23.

Pour $u \in \mathbb{R}^N$, on note $u_1, \ldots, u_N$ les composantes de u. Pour $u \in \mathbb{R}^N$, on dit que $u \geq 0$ si $u_i \geq 0$ pour tout $i \in \{1, \ldots, N\}$. Pour $u, v \in \mathbb{R}^N$, on note $u \cdot v = \sum_{i=1}^N u_i v_i$.

On munit $\mathbb{R}^N$ de la norme suivante : pour $u \in \mathbb{R}^N$, $\|u\| = \max\{|u_i|, i \in \{1, \ldots, N\}\}$. On munit alors $\mathcal{M}_N(\mathbb{R})$ de la norme induite, également notée $\|\cdot\|$, c'est-à-dire $\|B\| = \max\{\|Bu\|,\ u \in \mathbb{R}^N \text{ t.q. } \|u\| = 1\}$, pour tout $B \in \mathcal{M}_N(\mathbb{R})$.

Partie I Conditionnement de la matrice et borne sur l'erreur relative

1. (Existence et positivité de A^{-1}) Soient $b \in \mathbb{R}^N$ et $u \in \mathbb{R}^N$ t.q. $Au = b$. Remarquer que $Au = b$ peut s'écrire :

   $$\left\{ \begin{array}{l} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i,\ \forall i \in \{1, \ldots, N\}, \\ u_0 = u_{N+1} = 0. \end{array} \right. \tag{1.6.55}$$

   Montrer que $b \geq 0 \Rightarrow u \geq 0$. [On pourra considèrer $p \in \{0, \ldots, N+1\}$ t.q. $u_p = \min\{u_j, j \in \{0, \ldots, N+1\}$.]

   En déduire que A est inversible.

2. (Préliminaire...) On considère la fonction $\varphi \in C([0,1],\mathbb{R})$ définie par $\varphi(x) = (1/2)x(1-x)$ pour tout $x \in [0,1]$. On définit alors $\phi \in \mathbb{R}^N$ par $\phi_i = \phi(ih)$ pour tout $i \in \{1,\dots,N\}$. Montrer que $(A\phi)_i = 1$ pour tout $i \in \{1,\dots,N\}$.
3. (calcul de $\|A^{-1}\|$) Soient $b \in \mathbb{R}^N$ et $u \in \mathbb{R}^N$ t.q. $Au = b$. Montrer que $\|u\| \leq (1/8)\|b\|$ [Calculer $A(u \pm \|b\|\phi)$ avec ϕ défini à la question 2 et utiliser la question 1]. En déduire que $\|A^{-1}\| \leq 1/8$ puis montrer que $\|A^{-1}\| = 1/8$.
4. (calcul de $\|A\|$) Montrer que $\|A\| = \frac{4}{h^2}$.
5. (Conditionnement pour la norme $\|\cdot\|$). Calculer $\|A^{-1}\|\|A\|$. Soient $b, \delta_b \in \mathbb{R}^N$. Soient $u, \delta_u \in \mathbb{R}^N$ t.q. $Au = b$ et $A(u+\delta_u) = b + \delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \|A^{-1}\|\|A\| \dfrac{\|\delta_b\|}{\|b\|}$.
   Montrer qu'un choix convenable de b et δ_b donne l'égalité dans l'inégalité précédente.

Partie II Borne réaliste sur l'erreur relative : Conditionnement "efficace"

On se donne maintenant $f \in C([0,1],\mathbb{R})$ et on suppose (pour simplifier...) que $f(x) > 0$ pour tout $x \in]0,1[$. On prend alors, dans cette partie, $b_i = f(ih)$ pour tout $i \in \{1,\dots,N\}$. On considère aussi le vecteur φ défini à la question 2 de la partie I.

1. Montrer que
$$h\sum_{i=1}^{N} b_i\varphi_i \to \int_0^1 f(x)\phi(x)dx \text{ quand } N \to \infty$$
   et que
$$\sum_{i=1}^{N} b_i\varphi_i > 0 \text{ pour tout } N \in \mathbb{N}\ *\,.$$
   En déduire qu'il existe $\alpha > 0$, ne dépendant que de f, t.q. $h\sum_{i=1}^{N} b_i\varphi_i \geq \alpha$ pour tout $N \in \mathbb{N}^\star$.
2. Soit $u \in \mathbb{R}^N$ t.q. $Au = b$. Montrer que $N\|u\| \geq \sum_{i=1}^{N} u_i = u \cdot A\varphi \geq \frac{\alpha}{h}$ (avec α donné à la question 1).
   Soit $\delta_b \in \mathbb{R}^N$ et $\delta_u \in \mathbb{R}^N$ t.q. $A(u+\delta_u) = b + \delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \dfrac{\|f\|_{L^\infty(]0,1[)}}{8\alpha} \dfrac{\|\delta_b\|}{\|b\|}$.
3. Comparer $\|A^{-1}\|\|A\|$ (question I.5) et $\dfrac{\|f\|_{L^\infty(]0,1[)}}{8\alpha}$ (question II.2) quand N est "grand" (ou quand $N \to \infty$).

Exercice 29 (Méthode itérative du "gradient à pas fixe" [6]) *Suggestions en page 57*

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive, $b \in \mathbb{R}^N$ et $\alpha \in \mathbb{R}$. Pour trouver la solution de $Ax = b$, on considère la méthode itérative suivante :
– Initialisation : $x^{(0)} \in \mathbb{R}^N$,
– Iterations : $x^{(n+1)} = x^{(n)} + \alpha(b - Ax^{(n)})$.

1. Pour quelles valeurs de α (en fonction des valeurs propres de A) la méthode est-elle convergente ?
2. Calculer α_0 (en fonction des valeurs propres de A) t.q. $\rho(Id - \alpha_0 A) = \min\{\rho(Id - \alpha A),\ \alpha \in \mathbb{R}\}$.

Exercice 30 (Non convergence de la méthode de Jacobi) *Suggestions en page 57.*

Soit $a \in \mathbb{R}$ et
$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$
Montrer que A est symétrique définie positive si et seulement si $-1/2 < a < 1$ et que la méthode de Jacobi converge si et seulement si $-1/2 < a < 1/2$.

Exercice 31 (Une matrice cyclique) *Suggestions en page 58, corrigé en page* **??**

Soit $\alpha \in \mathbb{R}$ et soit $A \in \mathcal{M}_4(\mathbb{R})$ la matrice définie par

$$A = \begin{pmatrix} \alpha & -1 & 0 & -1 \\ -1 & \alpha & -1 & 0 \\ 0 & -1 & \alpha & -1 \\ -1 & 0 & -1 & \alpha \end{pmatrix}$$

Cette matrice est dite cyclique : chaque ligne de la matrice peut être déduite de la précédente en décalant chaque coefficient d'une position.

1. Déterminer les valeurs propres de A.
2. Pour quelles valeurs de α la matrice A est-elle symétrique définie positive ? singulière ?
3. On suppose ici que $\alpha \neq 0$. Soit $b = (b_1, b_2, b_3, b_4)^t \in \mathbb{R}^4$ donné. On considère la méthode de Jacobi pour la résolution du système $Ax = b$. Soit $(x^{(n)})_{n\in\mathbb{N}}$ la suite de vecteurs donnés par l'algorithme. On note $x_i^{(n)}$ pour $i = 1, \ldots, 4$ les composantes de $x^{(n)}$. Donner l'expression de $x_i^{(n+1)}$, $i = 1, \ldots, 4$, en fonction de $x_i^{(n)}$ et $b_i^{(n)}$, $i = 1, \ldots, 4$. Pour quelles valeurs de α la méthode de Jacobi converge-t-elle ?
4. On suppose maintenant que A est symétrique définie positive. Reprendre la question précédente pour la méthode de Gauss-Seidel.

Exercice 32 (Jacobi pour les matrices à diagonale dominante stricte) *Suggestions en page 58, corrigé en page 75*

Soit $A = (a_{i,j})_{i,j=1,\ldots,N} \in \mathcal{M}_N(\mathbb{R})$ une matrice à diagonale dominante stricte (c'est-à-dire $|a_{i,i}| > \sum_{j\neq i} |a_{i,j}|$ pour tout $i = 1, \ldots, N$). Montrer que A est inversible et que la méthode de Jacobi (pour calculer la solution de $Ax = b$) converge.

Exercice 33 (Jacobi pour pour un problème de diffusion)

Soit $f \in C([0,1])$; on considère le système linéaire $Ax = b$ issu de la discrétisation par différences finies de pas uniforme égal à $h = \frac{1}{N+1}$ du problème suivant :

$$\begin{cases} -u''(x) + \alpha u(x) = f(x),\ x \in [0,1], \\ u(0) = 0, u(1) = 1, \end{cases} \tag{1.6.56}$$

où $\alpha \geq 0$.

1. Donner l'expression de A et b.
2. Montrer que la méthode de Jacobi appliquée à la résolution de ce système converge (distinguer les cas $\alpha > 0$ et $\alpha = 0$).

Exercice 34 (Jacobi et diagonale dominance forte) *Corrigé en page 76*

On considère une matrice $A \in \mathcal{M}_N(\mathbb{R})$ inversible.

1. Montrer que si A est symétrique définie positive alors tous ses coefficients diagonaux sont strictement positifs. En déduire que la méthode de Jacobi est bien définie.
2. On suppose maintenant que la matrice diagonale extraite de A, notée D, est inversible. On suppose de plus que
$$\forall i = 1, \ldots, N, |a_{i,i}| \geq \sum_{j\neq i} |a_{i,j}| \text{ et } \exists i_0; |a_{i_0,i_0}| > \sum_{j\neq i_0} |a_{i,j}|.$$
(On dit que la matrice est à diagonale fortement dominante). Soit J la matrice d'itération de la méthode de Jacobi.

(a) Montrer que $\rho(J) \leq 1$.

(b) Montrer que si $Jx = \lambda x$ avec $|\lambda| = 1$, alors $x_i = \|x\|$, $\forall i = 1, \ldots, N$. En déduire que $x = 0$ et que la méthode de Jacobi converge.

(c) Retrouver ainsi le résultat de la question 2 de l'exercice 33.

Exercice 35 (Diagonalisation dans $\mathbb{R}$)

Soit E un espace vectoriel réel de dimension $N \in \mathbb{N}$ muni d'un produit scalaire, noté $(\cdot, \cdot)$. Soient T et S deux applications linéaires symétriques de E dans E (T symétrique signifie $(Tx, y) = (x, Ty)$ pour tous x, $y \in E$). On suppose que T est "définie positive" (c'est-à-dire $(Tx, x) > 0$ pour tout $x \in E \setminus \{0\}$).

1. Montrer que T est inversible. Pour $x, y \in E$, on pose $(x, y)_T = (Tx, y)$. Montrer que l'application $(x, y) \to (x, y)_T$ définit un nouveau produit scalaire sur E.
2. Montrer que $T^{-1}S$ est symétrique pour le produit scalaire défini à la question précédente. En déduire, avec le lemme 1.12 page 8, qu'il existe une base de E, notée $\{f_1, \ldots, f_N\}$, et il existe $\{\lambda_1, \ldots, \lambda_N\} \subset \mathbb{R}$ t.q. $T^{-1}Sf_i = \lambda_i f_i$ pour tout $i \in \{1, \ldots, N\}$ et t.q. $(Tf_i/f_j) = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, N\}$.

Exercice 36 (Méthode de Jacobi et relaxation) *Suggestions en page 58, corrigé en page 77*

Soit $N \geq 1$. Soit $A = (a_{i,j})_{i,j=1,\ldots,N} \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. On note D la partie diagonale de A, $-E$ la partie triangulaire inférieure de A et $-F$ la partie triangulaire supérieure de A, c'est-à-dire :

$$\begin{array}{l} D = (d_{i,j})_{i,j=1,\ldots,N},\ d_{i,j} = 0 \text{ si } i \neq j,\ d_{i,i} = a_{i,i}, \\ E = (e_{i,j})_{i,j=1,\ldots,N},\ e_{i,j} = 0 \text{ si } i \leq j,\ e_{i,j} = -a_{i,j} \text{ si } i > j, \\ F = (f_{i,j})_{i,j=1,\ldots,N},\ f_{i,j} = 0 \text{ si } i \geq j,\ f_{i,j} = -a_{i,j} \text{ si } i < j. \end{array}$$

Noter que $A = D - E - F$. Soit $b \in \mathbb{R}^N$. On cherche à calculer $x \in \mathbb{R}^N$ t.q. $Ax = b$. On suppose que D est définie positive (noter que A n'est pas forcément inversible). On s'intéresse ici à la méthode de Jacobi (par points), c'est–à–dire à la méthode itérative suivante :

Initialisation. $x^{(0)} \in \mathbb{R}^N$
Itérations. Pour $n \in \mathbb{N}$, $Dx^{(n+1)} = (E + F)x^{(n)} + b$.

On pose $J = D^{-1}(E + F)$.

1. Montrer, en donnant un exemple avec $N = 2$, que J peut ne pas être symétrique.
2. Montrer que J est diagonalisable dans $\mathbb{R}$ et, plus précisement, qu'il existe une base de $\mathbb{R}^N$, notée $\{f_1, \ldots, f_N\}$, et il existe $\{\mu_1, \ldots, \mu_N\} \subset \mathbb{R}$ t.q. $Jf_i = \mu_i f_i$ pour tout $i \in \{1, \ldots, N\}$ et t.q. $Df_i \cdot f_j = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, N\}$.

   En ordonnant les valeurs propres de J, on a donc $\mu_1 \leq \ldots \leq \mu_N$, on conserve cette notation dans la suite.
3. Montrer que la trace de J est nulle et en déduire que $\mu_1 \leq 0$ et $\mu_N \geq 0$.

   On suppose maintenant que A et $2D - A$ sont symétriques définies positives et on pose $x = A^{-1}b$.
4. Montrer que la méthode de Jacobi (par points) converge (c'est-à-dire $x^{(n)} \to x$ quand $n \to \infty$). [Utiliser un théorème du cours.]

   On se propose maintenant d'améliorer la convergence de la méthode par une technique de relaxation. Soit $\omega > 0$, on considère la méthode suivante :
   Initialisation. $x^{(0)} \in \mathbb{R}^N$
   Itérations. Pour $n \in \mathbb{N}$, $D\tilde{x}^{(n+1)} = (E + F)x^{(n)} + b$, $x^{(n+1)} = \omega\tilde{x}^{(n+1)} + (1 - \omega)x^{(n)}$.

5. Calculer les matrices M_ω (inversible) et N_ω telles que $M_\omega x^{(n+1)} = N_\omega x^{(n)} + b$ pour tout $n \in \mathbb{N}$, en fonction de ω, D et A. On note, dans la suite $J_\omega = (M_\omega)^{-1} N_\omega$.
6. On suppose dans cette question que $(2/\omega)D - A$ est symétrique définie positive. Montrer que la méthode converge (c'est-à-dire que $x^{(n)} \to x$ quand $n \to \infty$.)
7. Montrer que $(2/\omega)D - A$ est symétrique définie positive si et seulement si $\omega < 2/(1 - \mu_1)$.
8. Calculer les valeurs propres de J_ω en fonction de celles de J. En déduire, en fonction des μ_i, la valeur "optimale" de ω, c'est-à-dire la valeur de ω minimisant le rayon spectral de J_ω.

Exercice 37 (Méthodes de Jacobi et Gauss Seidel pour une matrice 3×3)

On considère la matrice $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ et le vecteur $b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$. Soit $x^{(0)}$ un vecteur de $\mathbb{R}^3$ donné.

1. *Méthode de Jacobi*

1.a Ecrire la méthode de Jacobi pour la résolution du système $Ax = b$, sous la forme $x^{(k+1)} = B_J x^{(k)} + c_J$.

1.b Déterminer le noyau de B_J et en donner une base.

1.c Calculer le rayon spectral de B_J et en déduire que la méthode de Jacobi converge.

1.d Calculer $x^{(1)}$ et $x^{(2)}$ pour les choix suivants de $x^{(0)}$:

$$(i)\, x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii) x^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}.$$

2. *Méthode de Gauss-Seidel*.

2.a Ecrire la méthode de Gauss-Seidel pour la résolution du système $Ax = b$, sous la forme $x^{(k+1)} = B_{GS} x^{(k)} + c_{GS}$.

2.b Déterminer le noyau de B_{GS}.

2.c Calculer le rayon spectral de B_{GS} et en déduire que la méthode de Gauss-Seidel converge.

2.d Comparer les rayons spectraux de B_{GS} et B_J et vérifier ainsi un résultat du cours.

2.d Calculer $x^{(1)}$ et $x^{(2)}$ pour les choix suivants de $x^{(0)}$:

$$(i)\ \ x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii)\ \ x^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

3. *Convergence en un nombre fini d'itérations*.

3.1 Soit α et β des réels. Soit $u^{(0)} \in R$ et $(u^{(k)})_{k \in \mathbb{N}}$ la suite réelle définie par $u^{(k+1)} = \alpha u^{(k)} + \beta$.

3.1.a Donner les valeurs de α et β pour lesquelles la suite $(u^{(k)})_{k \in \mathbb{N}}$ converge.

3.1.b On suppose que $\alpha \neq 0$, et que la suite $(u^{(k)})_{k \in \mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer que s'il existe $K \in N$ tel que $u_K = \overline{u}$, alors $u^{(k)} = \overline{u}$ pour tout $k \in \mathbb{N}$.

3.2 Soit $N > 1$, B une matrice réelle carrée d'ordre N et $b \in \mathbb{R}^N$. Soit $u^{(0)} \in \mathbb{R}^N$ et $(u^{(k)})_{k \in \mathbb{N}}$ la suite définie par $u^{(k+1)} = Bu^{(k)} + c$.

3.2.a Donner les conditions sur B et c pour que la suite $(u^{(k)})_{k \in \mathbb{N}}$ converge pour tout choix de $u_0 \in \mathbb{R}^N$.

3.2.b On suppose que la suite $(u^{(k)})_{k \in \mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer qu'on peut avoir $u^{(1)} = \overline{u}$ avec $u^{(0)} \neq \overline{u}$.

Exercice 38 (Jacobi et Gauss-Seidel pour une matrice tridiagonale)

Soit $A = (a_{i,j})_{i,j=1,\dots,N} \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée d'ordre N tridiagonale, c'est-à-dire telle que $a_{i,j} = 0$ si $|i-j| > 1$, et telle que la matrice diagonale $D = diag(a_{i,i})_{i=1,\dots,N}$ soit inversible. On note $A = D - E - F$ où $-E$ (resp. $-F$) est la partie triangulaire inférieure (resp. supérieure) de A, et on note J et G les matrices d'itération des méthodes de Jacobi et Gauss-Seidel associées à la matrice A.

1.a. Pour $\mu \in \mathbb{C}, \lambda \neq 0$ et $x \in \mathbb{C}^N$, on note

$$x_\mu = (x_1, \mu x_2, \dots, \mu^{k-1} x_k, \mu^{N-1} x_N)^t.$$

Montrer que si λ est valeur propre de J associée au vecteur propre x, alors x_μ vérifie $(\mu E + \frac{1}{\mu} F)x_\mu = \lambda D x_\mu$. En déduire que si $\lambda \neq 0$ est valeur propre de J alors λ^2 est valeur propre de G.

1.b Montrer que si λ^2 est valeur propre non nulle de G, alors λ est valeur propre de J .

2. Montrer que $\rho(G) = \rho(J)^2$. En déduire que lorsqu'elle converge, la méthode de Gauss-Seidel pour la résolution du système $Ax = b$ converge plus rapidement que la méthode de Jacobi.

3. Soit $\mathcal{L}_\omega$ la matrice d'itération de la méthode SOR associée à A. Montrer que λ est valeur propre de J si et seulement si ν_ω est valeur propre de $\mathcal{L}_\omega$, où $\nu_\omega = \mu_\omega^2$ et μ_ω vérifie $\mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0$.
En déduire que

$$\rho(\mathcal{L}_\omega) = \max_{\lambda \text{ valeur propre de } J} \{|\mu_\omega|; \mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0\}.$$

Exercice 39 (Méthode de Jacobi pour des matrices particulières) *Suggestions en page 58, corrigé en page 79*

On note $\mathcal{M}_N(\mathbb{R})$ l'ensemble des matrices carrées d'ordre N à coefficients réels, et Id la matrice identité dans $\mathcal{M}_N(\mathbb{R})$. Soit $A = [a_{i,j}]_{i,j=1,\dots,N} \in \mathcal{M}_N(\mathbb{R})$. On suppose que :

$$a_{i,j} \leq 0, \forall i, j = 1, \dots, N, i \neq j, \tag{1.6.57}$$

$$a_{i,i} > 0, \forall i = 1, \dots, N. \tag{1.6.58}$$

$$\sum_{i=1}^{N} a_{i,j} = 0, \forall j = 1, \dots, N. \tag{1.6.59}$$

Soit $\lambda \in \mathbb{R}_+^*$.

1. Pour $x \in \mathbb{R}^N$, on définit

$$\|x\|_A = \sum_{i=1}^{N} a_{i,i}|x_i|.$$

Montrer que $\|\cdot\|_A$ est une norme sur $\mathbb{R}^N$.

2. Montrer que la matrice $\lambda Id + A$ est inversible.
3. On considère le système linéaire suivant :

$$(\lambda Id + A)u = b \tag{1.6.60}$$

Montrer que la méthode de Jacobi pour la recherche de la solution de ce système définit une suite $(u^{(k)})_{k\in N} \subset \mathbb{R}^N$.

4. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ vérifie :

$$\|u^{(k+1)} - u^{(k)}\|_A \leq (\frac{1}{1+\alpha})^k \|u^{(1)} - u^{(0)}\|_A,$$

où $\alpha = \min_{i=1,\dots,N} a_{i,i}$.

5. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ est de Cauchy, et en déduire qu'elle converge vers la solution du système (1.6.60).

Exercice 40 (Une méthode itérative particulière)

Soient $\alpha_1, \ldots, \alpha_n$ des réels strictement positifs, et A la matrice $n \times n$ de coefficients $a_{i,j}$ définis par :

$$\begin{cases} a_{i,i} = 2 + \alpha_i \\ a_{i,i+1} = a_{i,i-1} = -1 \\ a_{i,j} = 0 \text{ pour tous les autres cas.} \end{cases}$$

Pour $\beta > 0$ on considère la méthode itérative $Mx^{(k+1)} = Nx^{(k)} + b$ avec $A = M - N$ et $N = \text{diag}(\beta - \alpha_i)$ (c.à.d $\beta - \alpha_i$ pour les coefficients diagonaux, et 0 pour tous les autres).

1. Soit $\lambda \in \mathbb{C}$ une valeur propre de la matrice $M^{-1}N$; montrer qu'il existe un vecteur $x \in \mathbb{C}^n$ non nul tel que $Nx \cdot \overline{x} = \lambda Mx \cdot \overline{x}$ (où $\overline{x}$ désigne le conjugué de x). En déduire que toutes les valeurs propres de la matrice $M^{-1}N$ sont réelles.

2. Montrer que le rayon spectral $\rho(M^{-1}N)$ de la matrice vérifie : $\rho(M^{-1}N) \leq \max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$

3. Déduire de la question 1. que si $\beta > \dfrac{\overline{\alpha}}{2}$, où $\overline{\alpha} = \max_{i=1,n} \alpha_i$, alors $\rho(M^{-1}N) < 1$, et donc que la méthode itérative converge.

4. Trouver le paramètre β minimisant $\max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$.
(On pourra d'abord montrer que pour tout $\beta > 0$, $|\beta - \alpha_i| \leq \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$ pour tout $i = 1, \ldots, n$, avec $\underline{\alpha} = \min_{i=1,\ldots,n} \alpha_i$ et $\bar{\alpha} = \max_{i=1,\ldots,n} \alpha_i$ et en déduire que $\max_{i=1,n} |\beta - \alpha_i| = \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$) .

Exercice 41 (Une matrice 3×3) *Suggestions en page 58, corrigé en page 81*

Soit $A \in M_3(\mathbb{R})$ définie par $A = Id - E - F$ avec

$$E = -\begin{pmatrix} 0 & 2 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ et } F = -\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

1. Montrer que A est inversible.
2. Soit $0 < \omega < 2$. Montrer que pour $(\frac{1}{\omega} Id - E)$ est inversible si et seulement si $\omega \neq \sqrt{2}/2$.

   Pour $0 < \omega < 2, \omega \neq \sqrt{2}/2$, on considère la méthode itérative (pour trouver la solution de $Ax = b$) suivante :

   $$(\frac{1}{\omega} Id - E)x^{n+1} = (F + \frac{1-\omega}{\omega} Id)x^n + b.$$

   Il s'agit donc de la "méthode I" du cours avec $B = \mathcal{L}_\omega = (\frac{1}{\omega} Id - E)^{-1}(F + \frac{1-\omega}{\omega} Id)$.
3. Calculer, en fonction de ω, les valeurs propres de $\mathcal{L}_\omega$ et son rayon spectral.
4. Pour quelles valeurs de ω la méthode est-elle convergente ? Déterminer $\omega_0 \in]0, 2[$ t.q. $\rho(\mathcal{L}_{\omega_0}) = \min\{\rho(\mathcal{L}_\omega), \omega \in]0, 2[, \omega \neq \sqrt{2}/2\}$.

Exercice 42 (Méthode des directions alternées)

Soit $N \in \mathbb{N}$ et $N \geq 1$, Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée d'ordre N symétrique inversible et $b \in \mathbb{R}^N$.
On cherche à calculer $u \in \mathbb{R}^N$, solution du système linéaire suivant :

$$Au = b, \tag{1.6.61}$$

On suppose connues des matrices X et $Y \in \mathcal{M}_N(\mathbb{R})$, symétriques. Soit $\alpha \in \mathbb{R}_+^*$, choisi tel que $X + \alpha Id$ et $Y + \alpha Id$ soient définies positives (où Id désigne la matrice identité d'ordre N) et $X + Y + \alpha Id = A$.

Soit $u^{(0)} \in \mathbb{R}^N$, on propose, pour résoudre (1.6.61), la méthode itérative suivante :

$$\begin{cases} (X + \alpha Id)u^{(k+1/2)} = -Yu^{(k)} + b, \\ (Y + \alpha Id)u^{(k+1)} = -Xu^{(k+1/2)} + b. \end{cases} \tag{1.6.62}$$

1. Montrer que la méthode itérative (1.6.62) définit bien une suite $(u^{(k)})_{k\in\mathbb{N}}$ et que cette suite converge vers la solution u de (1.1.1) si et seulement si

$$\rho\big((Y + \alpha Id)^{-1}X(X + \alpha Id)^{-1}Y\big) < 1.$$

   (On rappelle que pour toute matrice carrée d'ordre N, $\rho(M)$ désigne le rayon spectral de la matrice M.)

2. Montrer que si les matrices $(X + \frac{\alpha}{2}Id)$ et $(Y + \frac{\alpha}{2}Id)$ sont définies positives alors la méthode (1.6.62) converge. On pourra pour cela (mais ce n'est pas obligatoire) suivre la démarche suivante :

   (a) Montrer que

$$\rho\big((Y + \alpha Id)^{-1}X(X + \alpha Id)^{-1}Y\big) = \rho\big(X(X + \alpha Id)^{-1}Y(Y + \alpha Id)^{-1}\big).$$

   (On pourra utiliser l'exercice 6 page 37).

   (b) Montrer que

$$\rho\big(X(X + \alpha Id)^{-1}Y(Y + \alpha Id)^{-1}\big) \leq \rho\big(X(X + \alpha Id)^{-1}\big)\rho\big(Y(Y + \alpha Id)^{-1}\big).$$

   (c) Montrer que $\rho\big(X(X + \alpha Id)^{-1}\big) < 1$ si et seulement si la matrice $(X + \frac{\alpha}{2}Id)$ est définie positive.

   (d) Conclure.

3. Soit $f \in C([0, 1] \times [0, 1])$ et soit A la matrice carrée d'ordre $N = M \times M$ obtenue par discrétisation de l'équation $-\Delta u = f$ sur le carré $[0, 1] \times [0, 1]$ avec conditions aux limites de Dirichlet homogènes $u = 0$ sur $\partial\Omega$, par différences finies avec un pas uniforme $h = \frac{1}{M}$, et b le second membre associé.

   (a) Donner l'expression de A et b.

   (b) Proposer des choix de X, Y et α pour lesquelles la méthode itérative (1.6.62) converge dans ce cas et qui justifient l'appellation "méthode des directions alternées" qui lui est donnée.

Exercice 43 (Méthode de la puissance) *Suggestions en page 59, corrigé en page 82*

1. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Soit $\lambda_N \in \mathbb{R}$ valeur propre de A t.q. $|\lambda_N| = \rho(A)$ et soit $x^{(0)} \in \mathbb{R}^N$. On suppose que $-\lambda_N$ n'est pas une valeur propre de A et que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_N Id)$. On définit la suite $(x^{(n)})_{n\in\mathbb{N}}$ par $x^{(n+1)} = Ax^{(n)}$ pour $n \in \mathbb{N}$. Montrer que

   (a) $\frac{x^{(n)}}{(\lambda_N)^n} \to x$, quand $n \to \infty$, avec $x \neq 0$ et $Ax = \lambda_N x$.

   (b) $\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} \to \rho(A)$ quand $n \to \infty$.

   Cette méthode de calcul s'appelle "méthode de la puissance".

2. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^N$. Pour calculer x t.q. $Ax = b$, on considère la méthode itérative appelée "méthode I" en cours, et on suppose B symétrique. Montrer que, sauf cas particuliers à préciser,

   (a) $\frac{\|x^{(n+1)}-x\|}{\|x^{(n)}-x\|} \to \rho(B)$ quand $n \to \infty$ (ceci donne une estimation de la vitesse de convergence).

   (b) $\frac{\|x^{(n+1)}-x^{(n)}\|}{\|x^{(n)}-x^{(n-1)}\|} \to \rho(B)$ quand $n \to \infty$ (ceci permet d'estimer $\rho(B)$ au cours des itérations).

Exercice 44 (Méthode de la puissance inverse) *Suggestions en page 59.*

Soient $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique et $\lambda_1, \dots, \lambda_p$ ($p \leq N$) les valeurs propres de A. Soit $i \in \{1, \dots, p\}$, on cherche à calculer λ_i. Soit $x^{(0)} \in \mathbb{R}^N$. On suppose que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_i Id)$. On suppose également connaître $\mu \in \mathbb{R}$ t.q. $0 < |\mu - \lambda_i| < |\mu - \lambda_j|$ pour tout $j \neq i$. On définit la suite $(x^{(n)})_{n\in\mathbb{N}}$ par $(A - \mu Id)x^{(n+1)} = x^{(n)}$ pour $n \in \mathbb{N}$. Montrer que

1. $x^{(n)}(\lambda_i - \mu)^n \to x$, quand $n \to \infty$, avec $x \neq 0$ et $Ax = \lambda_i x$.
2. $\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} \to \frac{1}{|\mu-\lambda_i|}$ quand $n \to \infty$.

Exercice 45 (Méthode QR pour la recherche de valeurs propres)

Soient $\boldsymbol{u}$ et $\boldsymbol{v}$ deux vecteurs de $\mathbb{R}^N$. On rappelle que la projection orthogonale $\mathrm{proj}_{\boldsymbol{u}}(\boldsymbol{v})$ du vecteur $\boldsymbol{v}$ sur la droite vectorielle engendrée par $\boldsymbol{u}$ peut s'écrire de la manière suivante :

$$\mathrm{proj}_{\boldsymbol{u}}(\boldsymbol{v}) = \frac{\boldsymbol{v}\cdot\boldsymbol{u}}{\boldsymbol{u}\cdot\boldsymbol{u}}\boldsymbol{u},$$

où $\boldsymbol{u}\cdot\boldsymbol{v}$ désigne le produit scalaire des vecteurs $\boldsymbol{u}$ and $\boldsymbol{v}$. On note $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^N$.

1. Soient $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_N)$ une base de $\mathbb{R}^N$. On rappelle qu'à partir de cette base, on peut obtenir une base orthogonale $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_N)$ et une base orthonormale $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_N)$ par le procédé de Gram-Schmidt qu'on rappelle :

$$\begin{aligned}
&\boldsymbol{v}_1 = \boldsymbol{w}_1, && \boldsymbol{u}_1 = \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|}\\
&\boldsymbol{v}_2 = \boldsymbol{w}_2 - \mathrm{proj}_{\boldsymbol{v}_1}(\boldsymbol{w}_2), && \boldsymbol{u}_2 = \frac{\boldsymbol{v}_2}{\|\boldsymbol{v}_2\|}\\
&\boldsymbol{v}_3 = \boldsymbol{w}_3 - \mathrm{proj}_{\boldsymbol{v}_1}(\boldsymbol{w}_3) - \mathrm{proj}_{\boldsymbol{v}_2}(\boldsymbol{w}_3), && \boldsymbol{u}_3 = \frac{\boldsymbol{v}_3}{\|\boldsymbol{v}_3\|}\\
&\boldsymbol{v}_4 = \boldsymbol{w}_4 - \mathrm{proj}_{\boldsymbol{v}_1}(\boldsymbol{w}_4) - \mathrm{proj}_{\boldsymbol{v}_2}(\boldsymbol{w}_4) - \mathrm{proj}_{\boldsymbol{v}_3}(\boldsymbol{w}_4), && \boldsymbol{u}_4 = \frac{\boldsymbol{v}_4}{\|\boldsymbol{v}_4\|}\\
&\quad\vdots && \quad\vdots\\
&\boldsymbol{v}_k = \boldsymbol{w}_k - \sum_{j=1}^{k-1} \mathrm{proj}_{\boldsymbol{v}_j}(\boldsymbol{w}_k), && \boldsymbol{u}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}
\end{aligned}$$

On a donc

$$\boldsymbol{v}_k = \boldsymbol{w}_k - \sum_{j=1}^{k-1} \frac{\boldsymbol{w_k}\cdot\boldsymbol{v_j}}{\boldsymbol{v_j}\cdot\boldsymbol{v_j}}\boldsymbol{v_j}, \qquad \boldsymbol{u}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}. \tag{1.6.63}$$

1.a Montrer par récurrence que la famille $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_N)$ est une base orthogonale de $\mathbb{R}^N$.

1.b Soient A la matrice carrée d'ordre N dont les colonnes sont les vecteurs $\boldsymbol{w}_j$ et Q la matrice carrée d'ordre N dont les colonnes sont les vecteurs $\boldsymbol{u}_j$ définis par le procédé de Gram-Schmidt (1.6.63), ce qu'on note :

$$A = \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \dots & \boldsymbol{w}_N \end{bmatrix}, \qquad Q = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_N \end{bmatrix}.$$

Montrer que

$$\boldsymbol{w}_k = \|\boldsymbol{v}_k\|\boldsymbol{u}_k + \sum_{j=1}^{k-1} \frac{\boldsymbol{w_k} \cdot \boldsymbol{v_j}}{\|\boldsymbol{v_j}\|}\boldsymbol{u_j}.$$

En déduire que $A = QR$, où R est une matrice triangulaire supérieure dont les coefficients diagonaux sont positifs.

1.c. Montrer que pour toute matrice $A \in \mathcal{M}_N(\mathbb{R})$ inversible, on peut construire une matrice orthogonale Q (c.à. d. telle que $QQ^t = Id$) et une matrice triangulaire supérieure R à coefficients diagonaux positifs telles que $A = QR$.

1.d. Donner la décomposition QR de $A = \begin{bmatrix} 1 & 4 \\ 1 & 0 \end{bmatrix}$.

Soit A une matrice inversible.
Pour trouver les valeurs propres de A, on propose la méthode suivante, dite "méthode QR" : On pose $A_1 = A$ et on construit une matrice orthogonale Q_1 et une matrice triangulaire supérieure R_1 (par exemple construites à la question 2, bien que cette méthode ne soit pas celle utilisée en pratique, en raison d'instabilité numérique) telles que $A_1 = Q_1R_1$. On pose alors $A_2 = R_1Q_1$, qui est aussi une matrice inversible. On construit ensuite une matrice orthogonale Q_2 et une matrice triangulaire supérieure R_2 telles que $A_2 = Q_2R_2$ et on pose $A_3 = R_3Q_3$. On continue et on construit une suite de matrices A_k telles que :

$$A_1 = A = Q_1R_1,\ R_1Q_1 = A_2 = Q_2R_2,\ \dots,\ R_kQ_k = A_k = Q_{k+1}R_{k+1}. \tag{1.6.64}$$

Dans de nombreux cas, cette construction permet d'obenir les valeurs propres de la matrice A sur la diagonale des matrices A_k. Nous allons démontrer que ceci est vrai pour le cas particulier des matrices symétriques définies positives dont les valeurs propres sont simples (on peut le montrer pour une classe plus large de matrices).

On suppose à partir de maintenant que A est une matrice symétrique définie positive qui admet N valeurs propres (strictement positives) vérifiant $\lambda_1 > \lambda_2 > \ldots > \lambda_N$. On a donc :

$$A = P\Lambda P^t, \text{ avec } \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N), \text{ et } P \text{ est une matrice orthogonale.} \tag{1.6.65}$$

(La notation $\mathrm{diag}(\lambda_1, \dots, \lambda_N)$ désigne la matrice diagonale dont les termes diagonaux sont λ_1,... ,λ_N).

On suppose de plus que

$$P^t \text{ admet une décomposition } LU \text{ et que les coefficients diagonaux de } U \text{ sont strictement positifs.} \tag{1.6.66}$$

On va montrer que A_k tend vers $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$.

2. Soient Q_i et R_i les matrices orthogonales et triangulaires supérieures définies par (1.6.64).
2.1 Montrer que $A^2 = \tilde{Q}_2\tilde{R}_2$ avec $\tilde{Q}_k = Q_1Q_2$ et $\tilde{R}_k = R_2R_1$.

2.2 Montrer, par récurrence sur k, que

$$A^k = \tilde{Q}_k\tilde{R}_k, \tag{1.6.67}$$

avec

$$\tilde{Q}_k = Q_1Q_2 \dots Q_{k-1}Q_k \text{ et } \tilde{R}_k = R_kR_{k-1} \dots R_2R_1. \tag{1.6.68}$$

2.3 Justifier brièvement le fait que $\tilde{Q}_k$ est une matrice orthogonale et $\tilde{R}_k$ est une matrice triangulaire à coefficients diagonaux positifs.

3. Soit $M_k = \Lambda^k L \Lambda^{-k}$.

3.1 Montrer que $PM_k = \tilde{Q}_k T_k$ où $T_k = \tilde{R}_k U^{-1} \Lambda^{-k}$ est une matrice triangulaire supérieure dont les coefficients diagonaux sont positifs.

3.2 Calculer les coefficients de M_k en fonction de ceux de L et des valeurs propres de A.

3.3 En déduire que M_k tend vers la matrice identité et que $\tilde{Q}_k T_k$ tend vers P lorsque $k \to +\infty$.

4. Soient $(B_k)_{k\in\mathbb{N}}$ et $(C_k)_{k\in\mathbb{N}}$ deux suites de matrices telles que les matrices B_k sont orthogonales et les matrices C_k triangulaires supérieures et de coefficients diagonaux positifs. On va montrer que si $B_k C_k$ tend vers la matrice orthogonale B lorsque k tend vers l'infini alors B_k tend vers B et C_k tend vers l'identité lorsque k tend vers l'infini.
On suppose donc que $B_k C_k$ tend vers la matrice orthogonale B. On note $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N$ les colonnes de la matrice B et $\boldsymbol{b}_1^{(k)}, \boldsymbol{b}_2^{(k)}, \ldots, \boldsymbol{b}_N^{(k)}$ les colonnes de la matrice B_k, ou encore :

$$B = \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \ldots & \boldsymbol{b}_N \end{bmatrix}, \qquad B_k = \begin{bmatrix} \boldsymbol{b}_1^{(k)} & \boldsymbol{b}_2^{(k)} & \ldots & \boldsymbol{b}_N^{(k)} \end{bmatrix}.$$

et on note $c_{i,j}^{(k)}$ les coefficients de C_k.

4.1 Montrer que la première colonne de $B_k C_k$ est égale à $c_{1,1}^{(k)} \boldsymbol{b}_1^{(k)}$. En déduire que $c_{1,1}^{(k)} \to 1$ et que $\boldsymbol{b}_1^{(k)} \to \boldsymbol{b}_1$.

4.2 Montrer que la seconde colonne de $B_k C_k$ est égale à $c_{1,2}^{(k)} \boldsymbol{b}_1^{(k)} + c_{2,2}^{(k)} \boldsymbol{b}_2^{(k)}$. En déduire que $c_{1,2}^{(k)} \to 0$, puis que $c_{2,2}^{(k)} \to 1$ et que $\boldsymbol{b}_2^{(k)} \to \boldsymbol{b}_2$.

4.3 Montrer que lorsque $k \to +\infty$, on a $c_{i,j}^{(k)} \to 0$ si $i \neq j$, puis que $c_{i,i}^{(k)} \to 1$ et $\boldsymbol{b}_i^{(k)} \to \boldsymbol{b}_i$.

4.4 En déduire que B_k tend B et C_k tend vers l'identité lorsque k tend vers l'infini.

5. Déduire des questions 3 et 4 que $\tilde{Q}_k$ tend vers P et T_k tend vers Id lorsque $k \to +\infty$.

6. Montrer que $\tilde{R}_k (\tilde{R}_{k-1})^{-1} = T_k \Lambda T_{k-1}$. En déduire que R_k et A_k tendent vers Λ.

1.7 Suggestions

Exercice 1 page 36 (Matrices symétriques définies positives)

3. Utiliser la diagonalisation sur les opérateurs linéaires associés.

Exercice 3 page 36 (Normes induites particulières)

1. Pour montrer l'égalité, prendre x tel que $x_j = \mathrm{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\dots,N} |a_{i_0,j}| \geq \sum_{j=1,\dots,N} |a_{i,j}|$, $\forall i = 1,\dots,N$, et $\mathrm{sign}(s)$ désigne le signe de s.

2. Pour montrer l'égalité, prendre x tel que $x_{j_0} = 1$ et $x_j = 0$ si $j \neq j_0$, où j_0 est tel que $\sum_{i=1,\dots,N} |a_{i,j_0}| = \max_{j=1,\dots,N} \sum_{i=1,\dots,N} |a_{i,j}|$.

3. Utiliser le fait que $A^t A$ est une matrice symétrique positive pour montrer l'inégalité, et pour l'égalité, prendre pour x le vecteur propre associé à la plus grande valeur propre de A.

Exercice 5 page 37 (Rayon spectral)

1. Pour le sens direct, utiliser la proposition 1.7 page 6 du cours.

2. On rappelle que $\limsup_{k\to+\infty} u_k = \lim_{k\to+\infty} \sup_{n\geq k} u_n$, et $\liminf_{k\to+\infty} u_k = \lim_{k\to+\infty} \inf_{n\geq k} u_n$. Utiliser la question 1.

3. Utiliser le fait que $\liminf_{k\to+\infty} u_k$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$ (donc qu'il existe une suite extraite $(u_{k_n})_{n\in\mathbb{N}}$ telle que $uu_{k_n} \to \liminf_{k\to+\infty} u_k$ lorsque $k \to +\infty$.
4. Raisonner avec $\frac{1}{\alpha}A$ où $\alpha \in \mathbb{R}_+$ est tel que $\rho(A) < \alpha$ et utiliser la question 2 pour déduire que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq \rho(A).$$

Raisonner ensuite avec $\frac{1}{\beta}A$ où $\beta \in \mathbb{R}_+$ est tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$ et utiliser la question 3.

Exercice 9 page 37 (Série de Neumann)

1. Montrer que si $\rho(A) < 1$, alors 0 n'est pas valeur propre de $Id + A$ et $Id - A$.

2. Utiliser le résultat de la question 1 de l'exercice 5.

Exercice 16 page 39

2. Soit q le nombre de sur- ou sous-diagonales ($p = 2q + 1$). Compter le nombre c_q d'opérations nécessaires pour le calcul des colonnes 1 à q et $N - q + 1$ à N, puis le nombre d_n d'opérations nécessaires pour le calcul des colonnes $n = q + 1\ N - q$. En déduire l'estimation sur le nombre d'opérations nécessaires pour le calcul de toutes les colonnes, $Z_p(N)$, par :

$$2c_q \leq Z_p(N)2c_q + \sum_{n=q+1}^{N-q} c_n.$$

Exercice 18 page 40 (Propriétés générales du conditionnement)

Partie II
1. On rappelle que si A a comme valeurs propres $\lambda_1, \ldots, \lambda_N$, alors A^{-1} a comme valeurs propres $\lambda_1^{-1}, \ldots, \lambda_N^{-1}$ et A^t a comme valeurs propres $\lambda_1, \ldots, \lambda_N$.
2. Utiliser le fait que AA^t est diagonalisable.

5. Soient $0 < \lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$ et $0 < \mu_1 \leq \mu_2 \ldots \leq \mu_N$ les valeurs propres de A et B (qui sont s.d.p.). Montrer d'abord que :

$$cond_2(A+B) \leq \frac{\lambda_N + \mu_N}{\lambda_1 + \mu_1}.$$

Montrer ensuite que

$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}),\ \forall (a,b,c,d) \in (\mathbb{R}_+^*)^4.$$

et conclure

Exercice 26 page 44 (Valeurs propres et vecteurs propres de A.)

Chercher les vecteurs propres $\Phi \in \mathbb{R}^N$ de A sous la forme $\Phi_j = \varphi(x_j)$, $j = 1, \ldots, N$ où φ est introduite dans les indications de l'énoncé. Montrer que les valeurs propres associées à ces vecteurs propres sont de la forme : $\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos \frac{k\pi}{N+1})$.

Exercice 28 page 45 (Conditionnement efficace)

Partie 1
1. Pour montrer que A est inversible, utiliser le théorème du rang.

2. Utiliser le fait que φ est un polynôme de degré 2.

3. Pour montrer que $\|A^{-1}\| = \frac{1}{8}$, remarquer que le maximum de φ est atteint en $x = .5$, qui correspond à un point de discrétisation car N est impair.

Partie 2 Conditionnement efficace
1. Utiliser la convergence uniforme. 2. Utiliser le fait que $A\phi = (1 \ldots 1)^t$.

Exercice 29 page 46(Méthode itérative du "gradient à pas fixe".)

1. Calculer le rayon spectral $\rho(B)$ de la matrice d'itération $B = Id - \alpha A$. Calculer les valeurs de α pour lesquelles $\rho(B) < 1$ et en déduire que la méthode itérative du gradient à pas fixe converge si $0 < \alpha < \frac{2}{\rho(A)}$.
2. Remarquer que $\rho(Id - \alpha A) = \max(|1 - \alpha\lambda_1|, |1 - \alpha\lambda_N - 1|$, où $\lambda_1, \ldots, \lambda_N$ sont les valeurs propres de A ordonnées dans le sens croissant. En traçant les graphes des valeurs prises par $|1 - \alpha\lambda_1|$ et $|1 - \alpha\lambda_N - 1|$ en fonction de α, en déduire que le min est atteint pour $\alpha = \frac{2}{\lambda_1 + \lambda_N}$.

Exercice 30 page 46 (Non convergence de la méthode de Jacobi)

Considérer d'abord le cas $a = 0$.
Si $a \neq 0$, pour chercher les valeurs de a pour lesquelles A est symétrique définie positive, calculer les valeurs propres de A en cherchant les racines du polynôme caractéristique. Introduire la variable μ telle que $a\mu = 1 - \lambda$.

Pour chercher les valeurs de a pour lesquelles la méthode de Jacobi converge, calculer les valeurs propres de la matrice d'itération J définie en cours.

Exercice 31 page 47 (Une matrice cyclique)

1. On peut trouver les trois valeurs propres (dont une double) sans calcul en remarquant que pour $\alpha = 0$ il y a 2 fois 2 lignes identiques, que la somme des colonnes est un vecteur constant et par le calcul de la trace.
2. Une matrice A est symétrique définie positive si et seulement si elle est diagonalisable et toutes ses valeurs propres sont strictement positives.
3. Appliquer le cours.

Exercice 32 page 47 (Jacobi et diagonale dominante stricte.)

Pour montrer que A est inversible, montrer que $Ax = 0$ si et seulement si $x = 0$. Pour montrer que la méthode de Jacobi converge, montrer que toutes les valeurs propres de la matrice A sont strictement inférieures à 1 en valeur absolue.

Exercice 36 page 48 (Méthode de Jacobi et relaxation.)

1. Prendre pour A une matrice (2,2) symétrique dont les éléments diagonaux sont différents l'un de l'autre.
2. Appliquer l'exercice 35 page 48 en prenant pour T l'application linéaire dont la matrice est D et pour S l'application linéaire dont la matrice est $E + F$.
4. Remarquer que $\rho(J) = \max(-\mu_1, \mu_N)$, et montrer que :
si $\mu_1 \leq -1$, alors $2D - A$ n'est pas définie positive,
si $\mu_N \geq 1$, alors A n'est pas définie positive.
6. Reprendre le même raisonnement qu'à la question 2 à 4 avec les matrices M_ω et N_ω au lieu de D et $E + F$.
7. Chercher une condition qui donne que toutes les valeurs propres sont strictement positives en utilisant la base de vecteurs propres ad hoc. (Utiliser la base de $\mathbb{R}^N$, notée $\{f_1, \ldots, f_N\}$, trouvée à la question 2.)
8. Remarquer que les f_i de la question 2 sont aussi vecteurs propres de J_ω et en déduire que les valeurs propres $\mu_i^{(\omega)}$ de J_ω sont de la forme $\mu_i^{(\omega)} = \omega(\mu_i - 1 - 1/\omega)$. Pour trouver le paramètre optimal ω_0, tracer les graphes des fonctions de $\mathbb{R}_+$ dans $\mathbb{R}$ définies par $\omega \mapsto |\mu_1^{(\omega)}|$ et $\omega \mapsto |\mu_N^{(\omega)}|$, et en conclure que le minimum de $\max(|\mu_1^{(\omega)}|, |\mu_N^{(\omega)}|)$ est atteint pour $\omega = \frac{2}{2-\mu_1-\mu_N}$.

Exercice 39 page 50 (Méthode de Jacobi et relaxation.)

2. Utiliser l'exercice 32 page 47

Exercice 41 page 51 (Convergence de SOR.)

1. Calculer le déterminant de A.
2. Calculer le déterminant de $\frac{I}{d}\omega - E$.
3. Remarquer que les valeurs propres de $\mathcal{L}_\omega$ annulent $\det(\frac{1-\omega}{\omega} Id + F - \lambda(\frac{I}{d}\omega - E))$. Après calcul de ce déterminant, on trouve $\lambda_1 = 1 - \omega$, $\lambda_2 = \frac{1-\omega}{1+\sqrt{2}\omega}$, $\lambda_3 = \frac{1-\omega}{1-\sqrt{2}\omega}$.
Montrer que si $\omega < \sqrt{2}$, $\rho(\mathcal{L}_\omega) = |\lambda_3|$ et que $\rho(\mathcal{L}_\omega) = |\lambda_1|$ si $\omega \geq \sqrt{2}$.
4. Utiliser l'expression des valeurs propres pour montrer que la méthode converge si $\omega > \frac{2}{1+\sqrt{2}}$ et que le paramètre de relaxation optimal est $\omega_0 = 1$.

Exercice 43 page 52 (Méthode de la puissance pour calculer le rayon spectral de A.)

1. Décomposer x_0 sur une base de vecteurs propres orthonormée de A, et utiliser le fait que $-\lambda_N$ n'est pas valeur propre.
2. a/ Raisonner avec $y^{(n)} = x^{(n)} - x$ où x est la solution de $Ax = b$ et appliquer la question 1.
b/ Raisonner avec $y^{(n)} = x^{(n+1)} - x^{(n)}$.

Exercice 44 page 53 (Méthode de la puissance inverse)

Appliquer l'exercice précédent à la matrice $B = (A - \mu Id)^{-1}$.

1.8 Corrigés

Exercice 1 page 36 (Matrices symétriques définies positives)

1. Supposons qu'il existe un élément diagonal $a_{i,i}$ négatif. Alors $Ae_i \cdot e_i \leq 0$ ce qui contredit le fait que A est définie positive.

2. Soit $x \in \mathbb{R}^N$, décomposons x sur la base orthonormée $(f_i)_{i=1,N}$: $x = \sum_{i=1}^N x_i f_i$. On a donc :

$$Ax \cdot x = \sum_{i=1}^N \lambda_i x_i^2. \tag{1.8.69}$$

Montrons d'abord que si les valeurs propres sont strictement positives alors A est définie positive :
Supposons que $\lambda_i \geq 0, \forall i = 1, \ldots, N$. Alors pour $\forall x \in \mathbb{R}^N$, d'après (1.8.69), $Ax \cdot x \geq 0$ et la matrice A est positive.
Supposons maintenant que $\lambda_i \geq 0, \forall i = 1, \ldots, N$. Alors pour $\forall x \in \mathbb{R}^N$, toujours d'après (1.8.69), $(Ax \cdot x = 0) \Rightarrow (x = 0)$, et et la matrice A est donc bien définie.

Montrons maintenant la réciproque :
Si A est positive, alors $Af_i \cdot f_i \geq 0, \forall i = 1, \ldots, N$ et donc $\lambda_i \geq 0, \forall i = 1, \ldots, N$.
Si A est définie, alors $(A\alpha f_i \cdot \alpha f_i = 0) \Rightarrow (\alpha = 0), \forall i = 1, \ldots, N$ et donc $\lambda_i > 0, \forall i = 1, \ldots, N$.
3. Comme A est s.d.p., toutes ses valeurs propres sont strictement positives, et on peut donc définir l'application linéaire S dans la base orthonormée $(f_i)_{i=1,N}$ par : $S(f_i) = \sqrt{\lambda_i} f_i, \forall i = 1, \ldots, N$. On a évidemment $S \circ S = T$, et donc si on désigne par B la matrice représentative de l'application S dans la base canonique, on a bien $B^2 = A$.

Exercice 3 page 36 (Normes induites particulières)

1. Par définition, $\|A\|_\infty = \sup_{x \in \mathbb{R}^N, \|x\|_\infty = 1} \|Ax\|_\infty$, et

$$\|Ax\|_\infty = \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} a_{i,j} x_j| \leq \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}||x_j|.$$

Or $\|x\|_\infty = 1$ donc $|x_j| \leq 1$ et

$$\|Ax\|_\infty \leq \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}|.$$

Posons maintenant $\alpha = \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}|$ et montrons qu'il existe $x \in \mathbb{R}^N$, $\|x\|_\infty = 1$, tel que $\|Ax\|_\infty = \alpha$. Pour $s \in \mathbb{R}$, on note $\text{sign}(s)$ le signe de s, c'est–à–dire $\text{sign}(s) = s/|s|$ si $s \neq 0$ et $sign(0) = 0$. Choisissons $x \in \mathbb{R}^N$ défini par $x_j = \text{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\ldots,N} |a_{i_0,j}| \geq \sum_{j=1,\ldots,N} |a_{i,j}|, \forall i = 1, \ldots, N$. On a bien $\|x\|_\infty = 1$, et

$$\|Ax\|_\infty = \max_{i=1,\ldots,N} |\sum_{j=1}^N a_{i,j} sgn(a_{i_0,j})|.$$

Or, par choix de x, on a $\sum_{j=1,\ldots,N} |a_{i_0,j}| = max_{i=1,\ldots,N} \sum_{j=1,\ldots,N} |a_{i,j}|$. On en déduit que pour ce choix de x, on a bien $\|Ax\| = \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}|$.

2. Par définition, $\|A\|_1 = \sup_{x\in\mathbb{R}^N,\|x\|_1=1} \|Ax\|_1$, et

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1,\dots,N} |\sum_{j=1,\dots,N} a_{i,j}x_j| \leq \sum_{j=1,\dots,N} |x_j \left(|\sum_{i=1,\dots,N} |a_{i,j}| \right) \\ &\leq \max_{j=1,\dots,N} |\sum_{i=1,\dots,N} |a_{i,j}| \sum_{j=1,\dots,N} |x_j|.\end{aligned}$$

Et comme $\sum_{j=1,\dots,N} |x_j| = 1$, on a bien que $\|A\|_1 \leq \max_{j=1,\dots,N} \sum_{i=1,\dots,N} |a_{i,j}|$.
Montrons maintenant qu'il existe $x \in \mathbb{R}^N$, $\|x\|_1 = 1$, tel que $\|Ax\|_1 = \sum_{i=1,\dots,N} |a_{i,j}|$. Il suffit de considérer pour cela le vecteur $x \in \mathbb{R}^N$ défini par $x_{j_0} = 1$ et $x_j = 0$ si $j \neq j_0$, où j_0 est tel que $\sum_{i=1,\dots,N} |a_{i,j_0}| = \max_{j=1,\dots,N} \sum_{i=1,\dots,N} |a_{i,j}|$. On vérifie alors facilement qu'on a bien $\|Ax\|_1 = \max_{j=1,\dots,N} \sum_{i=1,\dots,N} |a_{i,j}|$.

3. Par définition de la norme 2, on a :

$$\|A\|_2^2 = \sup_{x\in\mathbb{R}^N,\|x\|_2=1} Ax \cdot Ax = \sup_{x\in\mathbb{R}^N,\|x\|_2=1} A^tAx \cdot x.$$

Comme A^tA est une matrice symétrique positive (car $A^tAx \cdot x = Ax \cdot Ax \geq 0$), il existe une base orthonormée $(f_i)_{i=1,\dots,N}$ et des valeurs propres $(\mu_i)_{i=1,\dots,N}$, avec $0 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_N$ tels que $Af_i = \mu_i f_i$ pour tout $i \in \{1,\dots,N\}$. Soit $x = \sum_{i=1,\dots,N} \alpha_i f_i \in \mathbb{R}^N$. On a donc :

$$A^tAx \cdot x = \big(\sum_{i=1,\dots,N} \mu_i\alpha_i f_i\big) \cdot \big(\sum_{i=1,\dots,N} \alpha_i f_i\big) = \sum_{i=1,\dots,N} \alpha_i^2\mu_i \leq \mu_N\|x\|_2^2.$$

On en déduit que $\|A\|_2^2 \leq \rho(A^tA)$.
Pour montrer qu'on a égalité, il suffit de considérer le vecteur $x = f_N$; on a en effet $\|f_N\|_2 = 1$, et $\|Af_N\|_2^2 = A^tAf_N \cdot f_N = \mu_N = \rho(A^tA)$.

Exercice 5 page 37 (Rayon spectral)

1. Si $\rho(A) < 1$, grâce au résultat d'approximation du rayon spectral de la proposition 1.7 page 6, il existe $\varepsilon > 0$ tel que $\rho(A) < 1 - 2\varepsilon$ et une norme induite $\|.\|_{A,\varepsilon}$ tels que $\|A\|_{A,\varepsilon} = \mu \leq \rho(A) + \varepsilon = 1 - \varepsilon < 1$. Comme $\|.\|_{A,\varepsilon}$ est une norme matricielle, on a $\|A^k\|_{A,\varepsilon} \leq \mu^k \to 0$ lorsque $k \to \infty$. Comme l'espace $\mathcal{M}_N(\mathbb{R})$ est de dimension finie, toutes les normes sont équivalentes, et on a donc $\|A^k\| \to 0$ lorsque $k \to \infty$.
Montrons maintenant la réciproque : supposons que $A^k \to 0$ lorsque $k \to \infty$, et montrons que $\rho(A) < 1$. Soient λ une valeur propre de A et x un vecteur propre associé. Alors $A^kx = \lambda^k x$, et si $A^k \to 0$, alors $A^kx \to 0$, et donc $\lambda^k x \to 0$, ce qui n'est possible que si $|\lambda| < 1$.

2. Si $\rho(A) < 1$, d'après la question précédente on a : $\|A^k\| \to 0$ donc il existe $K \in \mathbb{N}$ tel que pour $k \geq K$, $\|A^k\| < 1$. On en déduit que pour $k \geq K$, $\|A^k\|^{1/k} < 1$, et donc en passant à la limite sup sur k, $\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq 1$.

3. Comme $\liminf_{k\to+\infty} \|A^k\|^{1/k} < 1$, il existe une sous-suite $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ telle que $\|A^{k_n}\|^{1/k_n} \to \ell < 1$ lorsque $n \to +\infty$, et donc il existe N tel que pour $n \geq N$, $\|A^{k_n}\|^{1/k_n} \leq \eta$, avec $\eta \in]0,1[$. On en déduit que pour $n \geq N$, $\|A^{k_n}\| \leq \eta^{k_n}$, et donc que $A^{k_n} \to 0$ lorsque $n \to +\infty$. Soient λ une valeur propre de A et x un vecteur propre associé, on a : $A^{k_n}x = \lambda^{k_n}x$; on en déduit que $|\lambda| < 1$, et donc que $\rho(A) < 1$.

4. Soit $\alpha \in \mathbb{R}_+$ tel que $\rho(A) < \alpha$. Alors $\rho(\frac{1}{\alpha}A) < 1$, et donc par la question 2,

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \alpha,\ \forall\alpha > \rho(A).$$

En faisant tendre α vers $\rho(A)$, on obtient donc :

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq \rho(A). \tag{1.8.70}$$

Soit maintenant $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. On a alors $\liminf_{k\to+\infty} \|(\frac{1}{\beta}A)^k\|^{\frac{1}{k}} < 1$ et donc par la question 3, $\rho(\frac{1}{\beta}A) < 1$, donc $\rho(A) < \beta$ pour tout $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. En faisant tendre β vers $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}$, on obtient donc

$$\rho(A) \leq \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.8.71}$$

De (1.8.70) et (1.8.71), on déduit que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \rho(A). \tag{1.8.72}$$

5. Si $\|.\|$ est une norme matricielle, alors $\|A^k\| \leq \|A\|^k$ et donc d'après la question précédente, $\rho(A) \leq \|A\|$.

6. On a montré que $\rho(A) < 1$ si et seulement si $A^k \to 0$ et que donc, si $\rho(A) \geq 1$ alors $A^k \not\to 0$. On rappelle aussi que $A^k \to 0$ si et seulement si $A^k y \to 0, \quad \forall y \in \mathbb{R}^N$.

($\Rightarrow$) On démontre l'implication par contraposée. Si $\rho(A) \geq 1$ il existe $y \in \mathbb{R}^N$ tel que $A^k y \underset{k\to+\infty}{\not\longrightarrow} 0$.

($\Leftarrow$) Supposons maintenant que $\rho(B) < 1$ alors l'égalité (1.5.29) donne

$$x^{(k)} - x = B^k(x^{(0)} - x) \underset{k\to+\infty}{\to} 0$$

car $\rho(B) < 1$. Donc $x^{(k)} \underset{k\to+\infty}{\longrightarrow} x = A^{-1}b$. La méthode est bien convergente.

Exercice 7 page 37 (Rayon spectral)

Il suffit de prendre comme norme la norme définie par : $\|x\| = \sum_{i=1}^N \alpha_i^2$ où les $(\alpha_i)_{i=1,N}$ sont les composantes de x dans la base des vecteurs propres associés à A.

Pour montrer que ceci est faux dans le cas où A n'est pas diagonalisable, il suffit de prendre $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, on a alors $\rho(A) = 0$, et comme A est non nulle, $\|A\| \neq 0$.

Exercice 9 page 37 (Série de Neumann)

1. Si $\rho(A) < 1$, les valeurs propres de A sont toutes différentes de 1 et -1. Donc 0 n'est pas valeur propre des matrices $Id - A$ et $Id + A$, qui sont donc inversibles.

2. Suppposons que $\rho(A) < 1$. Il est facile de remarquer que

$$(\sum_{k=0}^{N} A^k)(Id - A) = Id - A^{N+1}. \tag{1.8.73}$$

Si $\rho(A) < 1$, d'après la question 1. de l'exercice 5 page 37, on a $A^k \to 0$ lorsque $k \to 0$. De plus, $Id - A$ est inversible. On peut donc passer à la limite dans (1.8.73) et on a donc $(Id - A)^{-1} = \sum_{k=0}^{+\infty} A^k$.
Remarquons de plus que la série de terme général A^k est absolument convergente pour une norme $\|\cdot\|_{A,\epsilon}$ donnée par la proposition 1.7 page 6, avec ε choisi tel que $\rho(A) + \varepsilon < 1$. Par contre, la série n'est pas absolument

convergente pour n'importe quelle norme. On pourra s'en convaincre facilement grâce au contre–exemple (en dimension 1) suivant : la série $s_k = 1 + x + \cdots + x^k$ est absolument convergente pour la norme $|\cdot|$ sur $\mathbb{R}$ pour $|x| < 1$, ce qui n'est évidemment plus le cas si l'on remplace la norme par la norme (pourtant équivalente) $\|\cdot\| = 10|\cdot|$.
Réciproquement, si $\rho(A) \geq 1$, la série ne peut pas converger en raison du résultat de la question 1 de l'exercice 5 page 37.

Exercice 11 page 38 (Décompositions LL^t et LDL^t)

1. On pose $L = \begin{pmatrix} 1 & 0 \\ \gamma & 1 \end{pmatrix}$ et $D = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$.
Par identification, on obtient $\alpha = 2$, $\beta = -\frac{1}{2}$ et $\gamma = \frac{1}{2}$.

Si maintenant on essaye d'écrire $A = LL^t$ avec $L = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$, on obtient $c^2 = -\frac{1}{2}$ ce qui est impossible dans $\mathbb{R}$.

En fait, on peut remarquer qu'il est normal que A n'admette pas de décomposition LL^t, car elle n'est pas définie positive. En effet, soit $x = (x_1, x_2)^t \in \mathbb{R}^2$,, alors $Ax \cdot x = 2x_1(x_1 + x_2)$, et en prenant $x = (1, -2)^t$, on a $Ax \cdot x < 0$.

2. 2. Reprenons en l'adaptant la démonstration du théorème 1.3. On raisonne donc par récurrence sur la dimension.

1. Dans le cas $N = 1$, on a $A = (a_{1,1})$. On peut donc définir $L = (\ell_{1,1})$ où $\ell_{1,1} = 1$, $D = (a_{1,1})$, $d_{1,1} \neq 0$, et on a bien $A = LDL^t$.
2. On suppose que, pour $1 \leq p \leq N$, la décomposition $A = LDL^t$ s'obtient pour $A \in \mathcal{M}_p(\mathbb{R})$ symétrique définie positive ou négative, avec $d_{i,i} \neq 0$ pour $1 \leq i \leq N$ et on va démontrer que la propriété est encore vraie pour $A \in \mathcal{M}_{N+1}(\mathbb{R})$ symétrique définie positive ou négative. Soit donc $A \in \mathcal{M}_{N+1}(\mathbb{R})$ symétrique définie positive ou négative ; on peut écrire A sous la forme :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.8.74}$$

où $B \in \mathcal{M}_N(\mathbb{R})$ est symétrique définie positive ou négative (calculer $Ax \cdot x$ avec $x = (y, 0)^t$, avec $y \in \mathbb{R}^N$ pour le vérifier), $a \in \mathbb{R}^N$ et $\alpha \in \mathbb{R}$.
Par hypothèse de récurrence, il existe une matrice $M \in \mathcal{M}_N(\mathbb{R})$ $M = (m_{i,j})_{i,j=1}^N$ et une matrice diagonale $\tilde{D} = \mathrm{diag}(d_{1,1}, d_{2,2}, \ldots, d_{N,N})$ dont les coefficients sont tous non nuls, telles que :
   (a) $m_{i,j} = 0$ si $j > i$
   (b) $m_{i,i} = 1$
   (c) $B = M\tilde{D}M^t$.
On va chercher L et D sous la forme :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right], \; D = \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right], \tag{1.8.75}$$

avec $b \in \mathbb{R}^N$, $\lambda \in \mathbb{R}$ tels que $LDL^t = A$. Pour déterminer b et λ, calculons LDL^t avec L et D de la forme (1.8.75) et identifions avec A :

$$LDL^t = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right] \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right] \left[\begin{array}{c|c} M^t & b \\ \hline 0 & 1 \end{array}\right] = \left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b+\lambda \end{array}\right]$$

On cherche $b \in \mathbb{R}^N$ et $\lambda \in \mathbb{R}$ tels que $LDL^t = A$, et on veut donc que les égalités suivantes soient vérifiées :

$$M\tilde{D}b = a \text{ et } b^t\tilde{D}b + \lambda = \alpha.$$

La matrice M est inversible (en effet, le déterminant de M s'écrit $det(M) = \prod_{i=1}^{N} 1 = 1$). Par hypothèse de récurrence, la matrice $\tilde{D}$ est aussi inversible. La première égalité ci-dessus donne : $b = \tilde{D}^{-1}M^{-1}a$. On calcule alors $\lambda = \alpha - b^tM^{-1}a$. Remarquons qu'on a forcément $\lambda \neq 0$, car si $\lambda = 0$,

$$A = LDL^t = \left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right]$$

qui n'est pas inversible. En effet, si on cherche $(x, y) \in \mathbb{R}^N \times \mathbb{R}$ solution de

$$\left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right] \left[\begin{array}{c} x \\ \hline y \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 0 \end{array}\right],$$

on se rend compte facilement que tous les couples de la forme $(-M^{-t}by, y)^t$, $y \in \mathbb{R}$, sont solutions. Le noyau de la matrice n'est donc pas réduit à $\{0\}$ et la matrice n'est donc pas inversible. On a ainsi montré que $d_{N+1,N+1} \neq 0$ ce qui termine la récurrence.

3. On reprend l'algorithme de décomposition LL^t :
Soit $A \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive ou négative ; on vient de montrer qu'il existe une matrice $L \in \mathcal{M}_N(\mathbb{R})$ triangulaire inférieure telle que $\ell_{i,j} = 0$ si $j > i$, $\ell_{i,i} = 1$, et une matrice $D \in \mathcal{M}_N(\mathbb{R})$ diagonale inversible, telles que et $A = LDL^t$. On a donc :

$$a_{i,j} = \sum_{k=1}^{N} \ell_{i,k}d_{k,k}\ell_{j,k}, \quad \forall\,(i,j) \in \{1,\dots,N\}^2. \tag{1.8.76}$$

1. Calculons la 1ère colonne de L ; pour $j = 1$, on a :

$$\begin{aligned} &a_{1,1} = d_{1,1} \text{ donc } d_{1,1} = a_{1,1}, \\ &a_{2,1} = \ell_{2,1}d_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{d_{1,1}}, \\ &a_{i,1} = \ell_{i,1}\ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{\ell_{1,1}} \ \forall i \in \{2,\dots,N\}. \end{aligned}$$

2. On suppose avoir calculé les n premières colonnes de L. On calcule la colonne $(n+1)$ en prenant $j = n+1$ dans (1.3.13)

   Pour $i = n+1$, $a_{n+1,n+1} = \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k} + d_{n+1,n+1}$ donc

$$d_{n+1,n+1} = a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k}. \tag{1.8.77}$$

On procède de la même manière pour $i = n+2, \ldots, N$; on a :

$$a_{i,n+1} = \sum_{k=1}^{n+1} \ell_{i,k} d_{k,k} \ell_{n+1,k} = \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} + \ell_{i,n+1} d_{n+1,n+1} \ell_{n+1,n+1}$$

et donc, comme on a montré dans la question 2 que les coefficients $d_{k,k}$ sont tous non nuls, on peut écrire :

$$\ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} \right) \frac{1}{d_{n+1,n+1}}. \tag{1.8.78}$$

Exercice 13 page 38 (Sur la méthode LL^t)

Corrigé en cours de rédaction

Exercice 14 page 38 (Sur la méthode LL^t)

Calculons le nombre d'opérations élémentaires nécessaires pour chacune des méthodes :

1. Le calcul de chaque coefficient nécessite N multiplications et $N-1$ additions, et la matrice comporte N^2 coefficients. Comme la matrice est symétrique, seuls $N(N+1)/2$ coefficients doivent être calculés. Le calcul de A^2 nécessite nécessite donc $\frac{(2N-1)N(N+1)}{2}$ opérations élémentaires.

   Le nombre d'opérations élémentaires pour effectuer la décomposition LL^t de A^2 nécessite $\frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6}$ (cours).

   La résolution du système $A^2 x = b$ nécessite $2N^2$ opérations (N^2 pour la descente, N^2 pour la remontée, voir cours).

   Le nombre total d'opérations pour le calcul de la solution du système $A^2 x = b$ par la première méthode est donc $\frac{(2N-1)N(N+1)}{2} + \frac{N^3}{3} + \frac{3N^2}{2} + \frac{N}{6} = \frac{4N^3}{3} + O(N^2)$ opérations.
2. La décomposition LL^t de A nécessite $\frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6}$, et la résolution des systèmes $LL^t y = b$ et $LL^t x = y$ nécessite $4N^2$ opérations. Le nombre total d'opérations pour le calcul de la solution du système $A^2 x = b$ par la deuxième méthode est donc $\frac{N^3}{3} + \frac{9N^2}{2} + \frac{N}{6} = \frac{N^3}{3} + O(N^2)$ opérations.

Pour les valeurs de N assez grandes, il est donc avantageux de choisir la deuxième méthode.

Exercice 16 page 39 (Décomposition LL^t d'une matrice bande)

On utilise le résultat de conservation du profil de la matrice énoncé dans le cours. Comme A est symétrique, le nombre p de diagonales de la matrice A est forcément impair si A ; notons $q = \frac{p-1}{2}$ le nombre de sous- et sur-diagonales non nulles de la matrice A, alors la matrice L aura également q sous-diagonales non nulles.

1. Cas d'une matrice tridiagonale. Si on reprend l'algorithme de construction de la matrice L vu en cours, on remarque que pour le calcul de la colonne $n+1$, avec $1 \leq n < N-1$, on a le nombre d'opérations suivant :

– Calcul de $\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k} \ell_{n+1,k})^{1/2} > 0$:
une multiplication, une soustraction, une extraction de racine, soit 3 opérations élémentaires.

– Calcul de $\ell_{n+2,n+1} = \left(a_{n+2,n+1} - \sum_{k=1}^{n} \ell_{n+2,k} \ell_{n+1,k} \right) \frac{1}{\ell_{n+1,n+1}}$:
une division seulement car $\ell_{n+2,k} = 0$.

On en déduit que le nombre d'opérations élémentaires pour le calcul de la colonne $n+1$, avec $1 \leq n < N-1$, est de 4.
Or le nombre d'opérations pour la première et dernière colonnes est inférieur à 4 (2 opérations pour la première colonne, une seule pour la dernière). Le nombre $Z_1(N)$ d'opérations élémentaires pour la décomposition LL^t de A peut donc être estimé par : $4(N-2) \leq Z_1(N) \leq 4N$, ce qui donne que $Z_1(N)$ est de l'ordre de $4N$ (le calcul exact du nombre d'opérations, inutile ici car on demande une estimation, est $4N-3$.)

2. Cas d'une matrice à p diagonales.
On cherche une estimation du nombre d'opérations $Z_p(N)$ pour une matrice à p diagonales non nulles (ou q sous-diagonales non nulles) en fonction de N.
On remarque que le nombre d'opérations nécessaires au calcul de

$$\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}\ell_{n+1,k})^{1/2} > 0, \tag{1.8.79}$$

$$\text{et } \ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k}\ell_{n+1,k} \right) \frac{1}{\ell_{n+1,n+1}}, \tag{1.8.80}$$

est toujours inférieur à $2q+1$, car la somme $\sum_{k=1}^{n}$ fait intervenir au plus q termes non nuls.
De plus, pour chaque colonne $n+1$, il y a au plus $q+1$ coefficients $\ell_{i,n+1}$ non nuls, donc au plus $q+1$ coefficients à calculer. Donc le nombre d'opérations pour chaque colonne peut être majoré par $(2q+1)(q+1)$.
On peut donc majorer le nombre d'opérations z_q pour les q premières colonnes et les q dernières par $2q(2q+1)(q+1)$, qui est indépendant de N (on rappelle qu'on cherche une estimation en fonction de N, et donc le nombre z_q est $O(1)$ par rapport à N.)
Calculons maintenant le nombre d'opérations x_n nécessaires une colonne $n = q+1$ à $N-q-1$. Dans (1.8.79) et (1.8.80), les termes non nuls de la somme sont pour $k = i-q, \ldots, n$, et donc on a $(n-i+q+1)$ multiplications et additions, une division ou extraction de racine. On a donc

$$\begin{aligned} x_n &= \sum_{i=n+1}^{n+q+1} \big(2(n-i+q+1)+1\big) \\ &= \sum_{j=1}^{q+1} \big(2(-j+q+1)+1\big) \\ &= (q+1)(2q+3) - 2\sum_{j=1}^{q+1} j \\ &= (q+1)^2. \end{aligned}$$

Le nombre z_i d'opérations nécessaires pour les colonnes $n = q+1$ à $N-q-1$ est donc

$$z_i = (q+1)^2(N-2q).$$

Un encadrement du nombre d'opérations nécessaires pour la décomposition LL^t d'une matrice à p diagonales est donc donnée par :

$$(q+1)^2(N-2q) \leq Z_p(N) \leq (q+1)^2(N-2q) + 2q(2q+1)(q+1), \tag{1.8.81}$$

et que, à q constant, $Z_p(N) = O((q+1)^2 N))$. Remarquons qu'on retrouve bien l'estimation obtenue pour $q = 1$.

3. Dans le cas de la discrétisation de l'équation $-u'' = f$ traitée dans le cours page 18, on a $q = 1$ et la méthode de Choleski nécessite de l'ordre de $4N$ opérations élémentaires, alors que dans le cas de la discrétisation de l'équation

$-\Delta u = f$ traitée dans le cours page 25-26, on a $q = \sqrt{N}$ et la méthode de Choleski nécessite de l'ordre de N^2 opérations élémentaires (dans les deux cas N est le nombre d'inconnues).

On peut noter que l'encadrement (1.8.81) est intéressant dès que q est d'ordre inférieur à N^α, $\alpha < 1$.

Exercice 18 page 40 (propriétés générales du conditionnement)

Partie I

1. Comme $\|\cdot\|$ est une norme induite, c'est donc une norme matricielle. On a donc pour toute matrice $A \in \mathcal{M}_N(\mathbb{R})$,

$$\|Id\| \leq \|A\| \ \|A^{-1}\|$$

ce qui prouve que $\mathrm{cond}(A) \geq 1$.

2.Par définition,

$$\begin{aligned} \mathrm{cond}(\alpha A) &= \|\alpha A\| \ \|(\alpha A)^{-1}\| \\ &= |\alpha| \ \|A\| \frac{1}{|\alpha|} \|A^{-1}\| = \mathrm{cond}(A) \end{aligned}$$

3. Soient A et B des matrices inversibles, alors AB est une matrice inversible et

$$\begin{aligned} \mathrm{cond}(AB) &= \|AB\| \ \|(AB)^{-1}\| = \|AB\| \ B^{-1}A^{-1}\| \\ &\leq \|A\| \ \|B\| \ \|B^{-1}\| \ \|A^{-1}\|, \end{aligned}$$

car $\|\cdot\|$ est une norme matricielle. Donc $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.

Partie II

1. Par définition, on a $\mathrm{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2$. Or on a vu à l'exercice 3 que $\|A\|_2 = (\rho(A^tA))^{1/2} = \sqrt{\sigma_N}$. On a donc

$$\|A^{-1}\|_2 = (\rho((A^{-1})^tA^{-1}))^{1/2} = \left(\rho(AA^t)^{-1})\right)^{1/2} \text{ ; or } \rho((AA^t)^{-1}) = \frac{1}{\tilde{\sigma}_1},$$

où $\tilde{\sigma}_1$ est la plus petite valeur propre de la matrice AA^t. Or les valeurs propres de AA^t sont les valeurs propres de A^tA : en effet, si λ est valeur propre de AA^t associée au vecteur propre x alors λ est valeur propre de AA^t associée au vecteur propre A^tx. On a donc

$$cond_2(A) = \sqrt{\frac{\sigma_N}{\sigma_1}}.$$

2. Si A est s.d.p., alors $A^tA = A^2$ et $\sigma_i = \lambda_i^2$ où λ_i est valeur propre de la matrice A. On a dans ce cas $cond_2(A) = \dfrac{\lambda_N}{\lambda_1}$.

3. Si $cond_2(A) = 1$, alors $\sqrt{\frac{\sigma_N}{\sigma_1}} = 1$ et donc toutes les valeurs propres de A^tA sont égales. Comme A^tA est symétrique définie positive (car A est inversible), il existe une base orthonormée $(f_1 \dots f_N)$ telle que $A^tAf_i = \sigma f_i, \ \forall i$ et $\sigma > 0$ (car A^tA est s.d.p.). On a donc $A^tA = \sigma Id \ A^t = \alpha^2 A^{-1}$ avec $\alpha = \sqrt{\sigma}$. En posant $Q = \dfrac{1}{\alpha}A$, on a donc $Q^t = \dfrac{1}{\alpha}A^t = \alpha A^{-1} = Q^{-1}$.

Réciproquement, si $A = \alpha Q$, alors $A^tA = \alpha^2 Id$, $\frac{\sigma_N}{\sigma_1} = 1$, et donc $cond_2(A) = 1$.

4. $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. On a donc :

$$cond_2(A) = \|A\|_2 \ \|A^{-1}\|_2 = \|QR\|_2 \ \|R^{-1}Q^t\|_2.$$

On a aussi $cond_2(A) = \sqrt{\frac{\sigma_N}{\sigma_1}}$ où $\sigma_1 \leq \ldots \leq \sigma_N$ sont les valeurs propres de A^tA. Or $A^tA = (QR)^t(QR) = R^tQ^{-1}QR = R^tR$. Donc $cond_2(A) = cond_2(R)$.

5. Soient $0 < \lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$ et $0 < \mu_1 \leq \mu_2 \ldots \leq \mu_N$ les valeurs propres de A et B (qui sont s.d.p.). Alors $cond_2(A+B) = \frac{\nu_N}{\nu_1}$, où $0 < \nu_1 \leq \ldots \leq \nu_N$ sont les valeurs propres de $A+B$.

a) On va d'abord montrer que

$$cond_2(A+B) \leq \frac{\lambda_N + \mu_N}{\lambda_1 + \mu_1}.$$

Remarquons en premier lieu que si A est s.d.p., alors

$$cond_2(A) = \frac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|_2=1} Ax \cdot x}$$

En effet, si A est s.d.p., alors $\sup_{\|x\|_2=1} Ax \cdot x = \lambda_N$; il suffit pour s'en rendre compte de décomposer x sur la base $(f_i)_{i=1\ldots N}$. Soit $x = \sum_{i=1}^{N} \alpha_i f_i$. Alors : $Ax \cdot x = \sum_{i=1}^{N} \alpha_i^2 \lambda_i \leq \lambda_N \Sigma \alpha_i^2 = \lambda_N$. Et $Af_N \cdot f_N = \lambda_N$.

De même, $Ax \cdot x \geq \lambda_1 \Sigma \alpha_i^2 = \lambda_1$ et $Ax \cdot x = \lambda_1$ si $x = f_1$. Donc $\inf_{\|x\|=1} Ax \cdot x = \lambda_1$.

On en déduit que si A est s.d.p., $cond_2(A) = \frac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|=1} Ax \cdot x}$

Donc $cond_2(A+B) = \frac{\sup_{\|x\|=1}(A+B)x \cdot x}{\inf_{\|x\|=1}(A+B)x \cdot x}$

$$\text{Or } \sup_{\|x\|=1} (Ax \cdot x + Bx \cdot x) \leq \sup_{\|x\|=1} Ax \cdot x + \sup_{\|x\|=1} Bx \cdot x = \lambda_N + \mu_N$$

$$\text{et } \inf_{\|x\|=1} (Ax \cdot x + Bx \cdot x) \geq \inf_{\|x\|=1} Ax \cdot x + \inf_{\|x\|=1} Bx \cdot x = \lambda_1 + \mu_1$$

donc

$$cond_2(A+B) \leq \frac{\lambda_N + \mu_N}{\lambda_1 + \mu_1}.$$

b) On va montrer que

$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}),\ \forall (a,b,c,d) \in (\mathbb{R}_+^*)^4.$$

Supposons que $\frac{a+b}{c+d} \geq \frac{a}{c}$ alors $(a+b)c \geq (c+d)a$ c'est–à–dire $bc \geq da$ donc $bc + bd \geq da + db$ soit $b(c+d) \geq d(a+b)$; donc $\frac{a+b}{c+d} \leq \frac{b}{d}$. On en déduit que $cond_2(A+B) \leq \max(cond_2(A), cond_2(B))$.

Exercice 20 page 40 (Minoration du conditionnement)

1. Comme A est inversible, $A + \delta A = A(Id + A^{-1}\delta A)$, et donc si $A + \delta A$ est singulière, alors $Id + A^{-1}\delta A$ est singulière. Or on a vu en cours que toute matrice de la forme $Id + B$ est inversible si et seulement si $\rho(B) < 1$. On en déduit que $\rho(A^{-1}\delta A) \geq 1$, et comme

$$\rho(A^{-1}\delta A) \leq \|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\|,$$

on obtient

$$\|A^{-1}\|\|\delta A\| \geq 1, \text{ soit encore } \mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta A\|}.$$

2. Soit $y \in \mathbb{R}^N$ tel que $\|y\| = 1$ et $\|A^{-1}y\| = \|A^{-1}\|$. Soit $x = A^{-1}y$, et $\delta A = \frac{-y\ x^t}{x^t\ x}$,on a donc

$$(A + \delta A)x = Ax - \frac{-y\ x^t}{x^t\ x}x = y - \frac{-y\ x^t x}{x^t\ x} = 0.$$

La matrice $A + \delta A$ est donc singulière. De plus,

$$\|\delta A\| = \frac{1}{\|x\|^2}\|y\ y^t\ A^{-t}\|.$$

Or par définition de x et y, on a $\|x\|^2 = \|A^{-1}\|^2$. D'autre part, comme il s'agit ici de la norme L^2, on a $\|A^{-t}\| = \|A^{-1}\|$. On en déduit que

$$\|\delta A\| = \frac{1}{\|A^{-1}\|^2}\|y\|^2\|A^{-1}\| = \frac{1}{\|A^{-1}\|}.$$

On a donc dans ce cas égalité dans (1.6.42).

3. Remarquons tout d'abord que la matrice A est inversible. En effet, $\det A = 2\alpha^2 > 0$. Soit $\delta A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\alpha & \alpha \\ 0 & -\alpha & -\alpha \end{pmatrix}$.

Comme $\det(A + \delta A) = 0$, la matrice $A + \delta A$ est singulière, et donc

$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta A\|}. \tag{1.8.82}$$

Or $\|\delta A\| = 2\alpha$ et $\|A\| = \max(3, 1 + 2\alpha) = 3$, car $\alpha \in]0, 1[$. Donc $\mathrm{cond}(A) \geq \frac{3}{2\alpha}$.

Exercice 22 page 41 (Calcul de l'inverse d'une matrice et conditionnement)

1. (a) L'inverse de la matrice A vérifie les quatre équations suivantes :

$$\begin{cases} X - A^{-1} = 0, & X^{-1} - A = 0, \\ AX - Id = 0, & XA - Id = 0. \end{cases}$$

Les quantités e_1, e_2, e_3 et e_4 sont les erreurs relatives commises sur ces quatre équations lorsqu'on remplace X par B ; en ce sens, elles mesurent la qualité de l'approximation de A^{-1}.

(b) On remarque d'abord que comme la norme est matricielle, on a $\|MP\| \leq \|M\|\|P\|$ pour toutes matrices M et P de $\mathcal{M}_N(\mathbb{R})$. On va se servir de cette propriété plusieurs fois par la suite.

(α) Comme $B = A^{-1} + E$, on a

$$e_1 = \frac{\|E\|}{\|A^{-1}\|} \leq \varepsilon\frac{\|A^{-1}\|}{\|A^{-1}\|} = \varepsilon.$$

(β) Par définition,

$$e_2 = \frac{\|B^{-1} - A\|}{\|A\|} = \frac{\|(A^{-1} + E)^{-1} - A\|}{\|A\|}.$$

Or

$$\begin{aligned} (A^{-1} + E)^{-1} - A &= (A^{-1}(Id + AE))^{-1} - A \\ &= (Id + AE)^{-1}A - A \\ &= (Id + AE)^{-1}(Id - (Id + AE))A \\ &= -(Id + AE)^{-1}AEA. \end{aligned}$$

On a donc

$$e_2 \leq \|(Id + AE)^{-1}\|\|A\|\|E\|.$$

Or par hypothèse, $\|AE\| \leq \|A\|\|E\| \leq \text{cond}(A)\varepsilon < 1$; on en déduit, en utilisant le théorème 1.11, que :

$$\|(Id + AE))^{-1}\| \leq \frac{1}{1 - \|AE\|}, \text{ et donc } e_2 \leq \frac{\varepsilon \text{cond}(A)}{1 - \varepsilon \text{cond}(A)}.$$

(γ) Par définition, $e_3 = \|AB - Id\| = \|A(A^{-1} + E) - Id\| = \|AE\| \leq \|A\|\|E\| \leq \|A\|\varepsilon\|A^{-1}\| = \varepsilon \text{cond}(A)$.

(δ) Enfin, $e_4 = \|BA - Id\| = \|(A^{-1} + E)A - Id\| \leq \|EA\| \leq \|E\|\|A\| \leq \varepsilon \text{cond}(A)$.

(c) (α) Comme $B = A^{-1}(Id + E')$, on a

$$e_1 = \frac{\|A^{-1}(Id + E') - A^{-1}\|}{\|A^{-1}\|} \leq \|Id + E' - Id\| \leq \varepsilon.$$

(β) Par définition,

$$\begin{aligned} e_2 &= \frac{\|(Id+E')^{-1}A - A\|}{\|A\|} \\ &= \frac{\|(Id+E')^{-1}(A-(Id+E')A)\|}{\|A\|} \\ &\leq \|(Id + E')^{-1}\|\|Id - (Id + E')\| \leq \frac{\varepsilon}{1-\varepsilon} \end{aligned}$$

car $\varepsilon < 1$ (théorème 1.1).

(γ) Par définition, $e_3 = \|AB - Id\| = \|AA^{-1}(Id + E') - Id\| = \|E'\| \leq \varepsilon$.

(δ) Enfin, $e_4 = \|BA - Id\| = \|A^{-1}(Id + E')A - Id\| = \|A^{-1}(A + E'A - A)\| \leq \|A^{-1}\|\|AE'\| \leq \varepsilon \text{cond}(A)$.

2. (a) On peut écrire $A + \delta_A = A(Id + A^{-1}\delta_A)$. On a vu en cours (théorème 1.11) que si $\|A^{-1}\delta_A\| < 1$, alors la matrice $Id + A^{-1}\delta_A$ est inversible. Or $\|A^{-1}\delta_A\| \leq \|A^{-1}\|\|\delta_A\|$, et donc la matrice $A + \delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.

(b) On peut écrire $\|(A + \delta_A)^{-1} - A^{-1}\| = \|(A + \delta_A)^{-1}(Id - (A + \delta_A)A^{-1}\| \leq \|(A + \delta_A)^{-1}\|\|Id - Id - \delta_A A^{-1}\| \leq \|(A + \delta_A)^{-1}\|\delta_A\|\|A^{-1}\|$. On en déduit le résultat.

Exercice 24 page 42 (IP-matrice)

1. Supposons d'abord que A est inversible et que $A^{-1} \geq 0$; soit $x \in \mathbb{R}^n$ tel que $b = Ax \geq 0$. On a donc $x = A^{-1}b$, et comme tous les coefficients de A^{-1} et de b sont positifs ou nuls, on a bien $x \geq 0$.
   Réciproquement, si A est une IP-matrice, alors $Ax = 0$ entraine $x = 0$ ce qui montre que A est inversible. Soit e_i le i-ème vecteur de la base canonique de $\mathbb{R}^n$, on a : $AA^{-1}e_i = e_i \geq 0$, et donc par la propriété de IP-matrice, $A^{-1}e_i \geq 0$, ce qui montre que tous les coefficients de A^{-1} sont positifs.

2. La matrice inverse de A est $A^{-1} = \frac{1}{\Delta}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ avec $\Delta = ad - bc$. Les coefficients de A^{-1} sont donc positifs ou nuls si et seulement si

$$\left\{ \begin{array}{l} ad < bc, \\ a \leq 0,\ d \leq 0 \\ b \geq 0,\ c \geq 0 \end{array} \right. \text{ ou } \left\{ \begin{array}{l} ad > bc, \\ a \geq 0,\ d \geq 0, \\ b \leq 0,\ c \leq 0. \end{array} \right.$$

   Or on a forcément $ad \neq 0$: en effet sinon on aurait dans le premier cas) $bc < 0$, or $b \leq 0$ et $c \leq 0$, ce qui aboutit à une contradiction. De même dans le deuxième cas, on aurait $bc > 0$, or $b \geq 0$ et $c \geq 0$. Les conditions précédentes sont donc équivalentes aux conditions (1.6.44).

3. La matrice A^t est une IP-matrice si et seulement A^t est inversible et $(A^t)^{-1} \geq 0$. Or $(A^t)^{-1} = (A^{-1})^t$. D'où l'équivalence.

4. Supposons que A vérifie (1.6.45), et soit $x \in \mathbb{R}^N$ tel que $Ax \geq 0$. Soit $k \in 1, \ldots, N$ tel que $x_k = \min\{x_i, i = 1, \ldots, N\}$. Alors
$$(Ax)_k = a_{k,k}x_k + \sum_{\substack{j=1 \\ j\neq k}}^{N} a_{k,j}x_j \geq 0.$$
Par hypothèse, $a_{k,j} \leq 0$ pour $k \neq j$, et donc $a_{k,j} = -|a_{k,j}|$. On peut donc écrire :
$$a_{k,k}x_k - \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{k,j}|x_j \geq 0,$$
et donc :
$$(a_{k,k} - \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{k,j}|)x_k \geq \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{k,j}|(x_j - x_k).$$
Comme $x_k = \min\{x_i, i = 1, \ldots, N\}$, on en déduit que le second membre de cette inégalité est positif ou nul, et donc que $x_k \geq 0$. On a donc $x \geq 0$.

5. Si la matrice A vérifie (1.6.46), alors la matrice A^t vérifie (1.6.45). On en déduit par les questions précédentes que A^t et A sont des IP-matrices.

6. Soit $\mathbf{1}$ le vecteur de $\mathbb{R}^N$ dont toutes les composantes sont égales à 1. Si $Ax > 0$, comme l'espace $\mathbb{R}^N$ est de dimension finie, il existe $\epsilon > 0$ tel que $Ax \geq \epsilon\mathbf{1}$. Soit $\eta = \epsilon A^{-1}\mathbf{1} \geq 0$; on a alors $A(x - \eta) \geq 0$ et donc $x \geq \eta$, car A est une IP-matrice.
Montrons maintenant que $\eta > 0$: tous les coefficients de A^{-1} sont positifs ou nuls et au moins l'un d'entre eux est non nul par ligne (puisque la matrice A^{-1} est inversible). On en déduit que $\eta_i = \epsilon \sum_{i=1}^{N} (A^{-1})_{i,j} > 0$ pour tout $i = 1, \ldots, N$. On a donc bien $x \geq \eta > 0$.

7. Soit A la matrice nulle, on a alors $\{x \in \mathbb{R}^N \text{ t.q. } Ax > 0\} = \emptyset$, et donc $\{x \in \mathbb{R}^N \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^N \text{ t.q. } x > 0\}$. Pourtant A n'est pas inversible, et n'est donc pas une IP-matrice.

8. Soit x tel que $Ax \geq 0$, alors il existe $\varepsilon \geq 0$ tel que $Ax + \varepsilon\mathbf{1} \geq 0$. Soit maintenant $b = A^{-1}\mathbf{1}$; on a $A(x + \varepsilon b) > 0$ et donc $x + \varepsilon b > 0$. En faisant tendre ε vers 0, on en déduit que $x \geq 0$.

9. Soit $T \in \mathcal{L}(E)$ défini par $f \in E \mapsto Tf$, avec $Tf(x) = f(\frac{1}{x})$ si $x \neq 0$ et $f(0) = \ell$, avec $\ell = \lim_{\pm\infty} f$. On vérifie facilement que $Tf \in E$. Si $Tf \geq 0$, alors $f(\frac{1}{x}) \geq 0$ pour tout $x \in \mathbb{R}$; donc $f(x) \geq 0$ pour tout $x \in \mathbb{R} \setminus \{0\}$; on en déduit que $f(0) \geq 0$ par continuité. On a donc bien $f \geq 0$.
Soit maintenant g définie de $\mathbb{R}$ dans $\mathbb{R}$ par $g(x) = |\arctan x|$. On a $g(0) = 0$, donc $g \not> 0$. Or $Tg(0) = \frac{\pi}{2}$ et $Tg(x) = |\arctan \frac{1}{x}| > 0$ si $x > 0$, donc $Tg > 0$.

Exercice 26 page 44 (Valeurs propres et vecteurs propres de A.)

1. Pour montrer que A est définie positive (car A est évidemment symétrique), on va montrer que $Ax \cdot x > 0$ si $x \neq 0$.
On a
$$Ax \cdot x = \frac{1}{h^2}\left[x_1(2x_1 - x_2) + \sum_{i=2}^{N-1} x_i(-x_{i-1} + 2x_i - x_{i+1}) + 2x_N^2 - x_{N-1}x_N\right]$$

On a donc

$$\begin{aligned} h^2 Ax \cdot x &= 2x_1^2 - x_1x_2 - \sum_{i=2}^{N-1} \left(x_i x_{i-1} + 2x_i^2\right) - \sum_{i=3}^{N} x_i x_{i-1} + 2x_N^2 - x_{N-1}x_N \\ &= \sum_{i=1}^{N} x_i^2 + \sum_{i=2}^{N} x_{1-i}^2 + x_N^2 - 2\sum_{i=1}^{N} x_i x_{i-1} \\ &= \sum_{i=2}^{N} (x_i - x_{i-1})^2 + x_1^2 + x_N^2 \geq 0. \end{aligned}$$

De plus, $Ax \cdot x = 0 \Rightarrow x_1^2 = x_N = 0$ et $x_i = x_{i-1}$ pour $i = 2$ à N, donc $x = 0$.
Pour chercher les valeurs propres et vecteurs propres de A, on s'inspire des valeurs propres et vecteurs propres du problème continu, c'est–à–dire des valeurs λ et fonctions φ telles que

$$\begin{cases} -\varphi''(x) = \lambda\varphi(x) & x \in]0,1[\\ \varphi(0) = \varphi(1) = 0 \end{cases} \tag{1.8.83}$$

(Notons que ce "truc" ne marche pas dans n'importe quel cas.)
L'ensemble des solutions de l'équation différentielle $-\varphi'' = \lambda\varphi$ est un espace vectoriel d'ordre 2, donc φ est de la forme $\varphi(x) = \alpha\cos\sqrt{\lambda}x + \beta\sin\sqrt{\lambda}x$ ($\lambda \geq 0$) et α et β dont déterminés par les conditions aux limites $\varphi(0) = \alpha = 0$ et $\varphi(1) = \alpha\cos\sqrt{\lambda} + \beta\sin\sqrt{\lambda} = 0$; on veut $\beta \neq 0$ car on cherche $\varphi \neq 0$ et donc on obtient $\lambda = k^2\pi^2$. Les couples (λ, φ) vérifiant (1.8.83) sont donc de la forme $(k^2\pi^2, \sin k\pi x)$.
2. Pour $k = 1$ à N, posons $\Phi_i^{(k)} = \sin k\pi x_i$, où $x_i = ih$, pour $i = 1$ à N, et calculons $A\Phi^{(k)}$:

$$(A\Phi^{(k)})_i = -sink\pi(i-1)h + 2\sin k\pi(ih) - \sin k\pi(i+1)h.$$

En utilisant le fait que $\sin(a+b) = \sin a\cos b + \cos a\sin b$ pour développer $\sin k\pi(1-i)h$ et $\sin k\pi(i+1)h$, on obtient (après calculs) :

$$(A\Phi^{(k)})_i = \lambda_k \Phi_i^{(k)}, \quad i = 1, \ldots, N,$$

où $\lambda_k = \dfrac{2}{h^2}(1 - \cos k\pi h) = \dfrac{2}{h^2}(1 - \cos\dfrac{k\pi}{N+1})$
On a donc trouvé N valeurs propres $\lambda_1 \ldots \lambda_N$ associées aux vecteurs propres $\Phi^{(1)} \ldots \Phi^{(N)}$ de $\mathbb{R}^N$ tels que $\Phi_i^{(k)} = \sin\dfrac{k\pi i}{N+1}$, $i = 1 \ldots N$.
Remarque : Lorsque $N \to +\infty$ (ou $h \to 0$), on a

$$\lambda_k^{(h)} = \frac{2}{h^2}\left(1 - 1 + \frac{k^2\pi^2h^2}{2} + O(h^4)\right) = k^2\phi^2 + O(h^2)$$

Donc

$$\begin{aligned} \lambda_k^{(h)} &\to k^2\pi^2 = \lambda_k \\ h &\to 0. \end{aligned}$$

Calculons $cond_2(A)$. Comme A est s.d.p., on a $cond_2(A) = \dfrac{\lambda_N}{\lambda_1} = \dfrac{1 - \cos\frac{N\pi}{N+1}}{1 - \cos\frac{\pi}{N+1}}$.
On a : $h^2\lambda_N = 2(1 - \cos\frac{N\pi}{N+1}) \to 4$ et $\lambda_1 \to \pi^2$ lorsque $h \to 0$. Donc $h^2 cond_2(A) \to \frac{4}{\pi^2}$ lorsque $h \to 0$.

Exercice 28 page 45 (Conditionnement “efficace")

Partie I

1. Soit $u = (u_1, \dots, u_N)^t$. On a

$$Au = b \Leftrightarrow \left\{ \begin{array}{l} \dfrac{1}{h^2}(u_i - u_{i-1}) + \dfrac{1}{h^2}(u_i - u_{i+1}) = b_i, \ \ \forall i = 1, \dots, N, \\ u_0 = u_{N+1} = 0. \end{array} \right.$$

Supposons $b_i \geq 0, \forall i = 1, \dots, N$, et soit $p \in \{0, \dots, N+1\}$ tel que $u_p = \min(u_i, \ \ i = 0, \dots, N+1)$.
Si $p = 0$ ou $N+1$, alors $u_i \geq 0 \ \forall i = 0, N+1$ et donc $u \geq 0$.
Si $p \in \{1, \dots, N\}$, alors

$$\frac{1}{h^2}(u_p - u_{p-1}) + \frac{1}{h^2}(u_p - u_{p+1}) \geq 0$$

et comme $u_p - u_{p-1} < 0$ et $u_p - u_{p+1} \leq 0$, on aboutit à une contradiction.

Montrons maintenant que A est inversible. On vient de montrer que si $Au \geq 0$ alors $u \geq 0$. On en déduit par linéarité que si $Au \leq 0$ alors $u \leq 0$, et donc que si $Au = 0$ alors $u = 0$. Ceci démontre que l'application linéaire représentée par la matrice A est injective donc bijective (car on est en dimension finie).

2. Soit $\varphi \in C([0,1], \mathbb{R})$ tel que $\varphi(x) = 1/2x(1-x)$ et $\phi_i = \varphi(x_i), \ i = 1, N$, où $x_i = ih$.
$(A\phi)_i$ est le développement de Taylor à l'ordre 2 de $\varphi''(x_i)$, et comme φ est un polynôme de degré 2, ce développement est exact. Donc $(A\phi)_i = \varphi''(x_i) = 1$.

3. Soient $b \in \mathbb{R}^N$ et $u \in \mathbb{R}^N$ tels que $Au = b$. On a :

$$(A(u \pm \|b\|\varphi))_i = (Au)_i \pm \|b\|(A\phi)_i = b_i \pm \|b\|.$$

Prenons d'abord $\tilde{b}_i = b_i + \|b\| \geq 0$, alors par la question (1),

$$u_i + \|b\|\phi_i \geq 0 \ \ \forall i = 1 \dots N.$$

Si maintenant on prend $\bar{b}_i = b_i - \|b\| \leq 0$, alors

$$u_i - \|b\|\phi_i \leq 0 \ \ \forall i = 1, \dots, N.$$

On a donc $-\|b\|\phi_i \leq \|b\|\phi_i$.
On en déduit que $\|u\|_\infty \leq \|b\| \ \ \|\phi\|_\infty$; or $\|\phi\|_\infty = \dfrac{1}{8}$. D'où $\|u\|_\infty \leq \dfrac{1}{8}\|b\|$.
On peut alors écrire que pour tout $b \in \mathbb{R}^N$,

$$\|A^{-1}b\|_\infty \leq \frac{1}{8}\|b\|, \text{ donc } \frac{\|A^{-1}b\|_\infty}{\|b\|_\infty} \leq \frac{1}{8}, \text{ d'où } \|A^{-1}\| \leq \frac{1}{8}.$$

On montre que $\|A^{-1}\| = \dfrac{1}{8}$ en prenant le vecteur b défini par $b(x_i) = 1, \forall i = 1, \dots, N$. On a en effet $A^{-1}b = \phi$, et comme N est impair, $\exists i \in \{1, \dots, N\}$ tel que $x_i = \frac{1}{2}$;or $\|\varphi\|_\infty = \varphi(\frac{1}{2}) = \frac{1}{8}$.

4. Par définition, on a $\|A\| = \sup_{\|x\|_\infty = 1} \|Ax\|$, et donc $\|A\| = \max_{i=1,N} \sum_{j=1,N} |a_{i,j}|$, d'où le résultat.

5. Grâce aux questions 3 et 4, on a, par définition du conditionnement pour la norme $\|\cdot\|$, $\mathrm{cond}(A) = \|A\|\|A^{-1}\| = \frac{1}{2h^2}$.

Comme $A\delta_u = \delta_b$, on a :

$$\|\delta_u\| \leq \|A^{-1}\|\|\delta_b\|\frac{\|b\|}{\|b\|} \leq \|A^{-1}\|\|\delta_b\|\frac{\|A\|\|u\|}{\|b\|},$$

d'où le résultat.
Pour obtenir l'égalité, il suffit de prendre $b = Au$ où u est tel que $\|u\| = 1$ et $\|Au\| = \|A\|$, et δ_b tel que $\|\delta_b\| = 1$ et $\|A^{-1}\delta_b\| = \|A^{-1}\|$. On obtient alors

$$\frac{\|\delta_b\|}{\|b\|} = \frac{1}{\|A\|} \text{ et } \frac{\|\delta u\|}{\|u\|} = \|A^{-1}\|.$$

D'où l'égalité.

Partie 2 Conditionnement efficace

1. Soient $\varphi^{(h)}$ et $f^{(h)}$ les fonctions constantes par morceaux définies par

$$\begin{array}{ll} \varphi^h(x) & = \left\{ \begin{array}{l} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[,\ i = 1, \ldots, N, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \text{ et} \\ f^{(h)}(x) & = \left\{ \begin{array}{l} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ \ x_i + \frac{h}{2}[, \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \end{array}$$

Comme $f \in C([0,1], \mathbb{R})$ et $\varphi \in C^2([0,1], \mathbb{R})$, la fonction f_h (resp. φ_h) converge uniformément vers f (resp. φ) lorsque $h \to 0$. On a donc

$$h\sum_{i=1}^{N} b_i\varphi_i = \int_0^1 f^{(h)}(x)\varphi^{(h)}(x)dx \to \int_0^1 f(x)\varphi(x)dx \text{ lorsque } h \to 0.$$

Comme $b_i > 0$ et $f_i > 0\ \forall i = 1, \ldots, N$, on a évidemment

$$S_N = \sum_{i=1}^{N} b_i\varphi_i > 0 \text{ et } S_N \to \int_0^1 f(x)\varphi(x)dx = \beta > 0 \text{ lorsque } h \to 0.$$

Donc il existe $N_0 \in \mathbb{N}$ tel que si $N \geq N_0$, $S_N \geq \frac{\beta}{2}$, et donc $S_N \geq \alpha = \min(S_0, S_1 \ldots S_{N_0}, \frac{\beta}{2}) > 0$.

2. On a $N\|u\| = N \sup_{i=1,N} |u_i| \geq \sum_{i=1}^{N} u_i$. D'autre part, $A\varphi = (1 \ldots 1)^t$ donc $u \cdot A\varphi = \sum_{i=1}^{N} u_i$; or $u \cdot A\varphi = A^t u \cdot \varphi = Au \cdot \varphi$ car A est symétrique. Donc $u \cdot A\varphi = \sum_{i=1}^{N} b_i\varphi_i \geq \frac{\alpha}{h}$ d'après la question 1. Comme $\delta_u = A^{-1}\delta_b$, on a donc $\|\delta_u\| \leq \|A^{-1}\|\ \|\delta_b\|$; et comme $N\|u\| \geq \frac{\alpha}{h}$, on obtient : $\frac{\|\delta_u\|}{\|u\|} \leq \frac{1}{8}\frac{hN}{\alpha}\|\delta_b\|\frac{\|f\|_\infty}{\|b\|}$. Or $hN = 1$ et on a donc bien :

$$\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|_\infty}{8\alpha}\frac{\|\delta_b\|}{\|b\|}.$$

3. Le conditionnement $\mathrm{cond}(A)$ calculé dans la partie 1 est d'ordre $1/h^2$, et donc tend vers l'infini lorsque le pas de discrétisation tend vers 0, alors qu'on vient de montrer dans la partie 2 que la variation relative $\frac{\|\delta_u\|}{\|u\|}$ est inférieure à une constante multipliée par la variation relative de $\frac{\|\delta_b\|}{\|b\|}$. Cette dernière information est nettement plus utile et réjouissante pour la résolution effective du système linéaire.

Exercice 30 page 46 (Non convergence de la méthode de Jacobi)

- Si $a = 0$, alors $A = Id$, donc A est s.d.p. et la méthode de Jacobi converge.
- Si $a \neq 0$, posons $a\mu = (1 - \lambda)$, et calculons le polynôme caractéristique de la matrice A en fonction de la variable μ.

$$P(\mu) = det \begin{vmatrix} a\mu & a & a \\ a & a\mu & a \\ a & a & a\mu \end{vmatrix} = a^3 det \begin{vmatrix} \mu & 1 & 1 \\ 1 & \mu & 1 \\ 1 & 1 & \mu \end{vmatrix} = a^3(\mu^3 - 3\mu + 2).$$

On a donc $P(\mu) = a^3(\mu - 1)^2(\mu + 2)$. Les valeurs propres de la matrice A sont donc obtenues pour $\mu = 1$ et $\mu = 2$, c'est–à–dire : $\lambda_1 = 1 - a$ et $\lambda_2 = 1 + 2a$.
La matrice A est définie positive si $\lambda_1 > 0$ et $\lambda_2 > 0$, c'est–à–dire si $-\frac{1}{2} < a < 1$.
La méthode de Jacobi s'écrit :

$$X^{(n+1)} = D^{-1}(D - A)X^{(n)},$$

avec $D = Id$ dans le cas présent ; donc la méthode converge si et seulement si $\rho(D - A) < 1$.
Les valeurs propres de $D - A$ sont de la forme $\nu = 1 - \lambda$ où λ est valeur propre de A. Les valeurs propres de $D - A$ sont donc $\nu_1 == -a$ (valeur propre double) et $\nu_2 = 2a$.. On en conclut que la méthode de Jacobi converge si et seulement si $-1 < -a < 1$ et $-1 < 2a < 1$, *i.e.* $\frac{1}{2} < a < \frac{1}{2}$.
La méthode de Jacobi ne converge donc que sur l'intervalle $]-\frac{1}{2}, \frac{1}{2}[$ qui est strictement inclus dans l'intervalle $]-\frac{1}{2}, 1[$ des valeurs de a pour lesquelles la matrice A est s.d.p..

Exercice 32 page 47 (Jacobi pour les matrices à diagonale dominante stricte)

Pour montrer que A est inversible, supposons qu'il existe $x \in \mathbb{R}^N$ tel que $Ax = 0$; on a donc

$$\sum_{j=1}^{N} a_{ij}x_j = 0.$$

Pour $i \in \{1, \ldots, N\}$, on a donc

$$|a_{i,i}||x_i| = |a_{i,i}x_i| = |\sum_{j;i\neq j} a_{i,j}x_j| \leq \sum_{j;i\neq j} |a_{i,j}|\|x\|_\infty, \quad \forall i = 1, \ldots, N.$$

Si $x \neq 0$, on a donc

$$|x_i| \leq \frac{\sum_{j;i\neq j} a_{i,j}x_j|}{|a_{i,i}|}\|x\|_\infty < \|x\|_\infty, \quad \forall i = 1, \ldots, N$$

, ce qui est impossible pour i tel que

$$|x_i| = \|x\|_\infty.$$

Montrons maintenant que la méthode de Jacobi converge : Avec le formalisme de la méthode II du cours, on a

$$M = D = \begin{bmatrix} a_{1,1} & & 0 \\ & \ddots & \\ 0 & & a_{N,N} \end{bmatrix}, \text{ et } N = M - A.$$

La matrice d'itération est

$$J = M^{-1}N = D^{-1}N = \begin{bmatrix} a_{1,1}^{-1} & & 0 \\ & \ddots & \\ 0 & & a_{N,N}^{-1} \end{bmatrix} \begin{bmatrix} 0 & & -a_{i,j} \\ & \ddots & \\ -a_{i,j} & & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \dots \\ & \ddots & \\ -\frac{a_{1,1}}{a_{N,N}} & \dots & 0 \end{bmatrix}.$$

Cherchons le rayon spectral de J : soient $x \in \mathbb{R}^N$ et $\lambda \in \mathbb{R}$ tels que $Jx = \lambda x$, alors

$$\sum_{j;i\neq j} -\frac{a_{i,j}}{a_{i,i}} x_j = \lambda x_i, \text{ et donc } |\lambda||x_i| \leq \sum_{j;i\neq j} |a_{i,j}| \frac{\|x\|_\infty}{|a_{i,i}|}.$$

Soit i tel que $|x_i| = \|x\|_\infty$ et $x \neq 0$, on déduit de l'inégalité précédente que $|\lambda| \leq \frac{\sum_{j;i\neq j} |a_{i,j}|}{|a_{ii}|} < 1$ pour toute valeur propre λ. On a donc $\rho(J) < 1$. Donc la méthode de Jacobi converge.

Exercice 34 page 47 (Jacobi pour les matrices à diagonale dominante forte)

1. Si A est symétrique définie positive alors $Ae_i \cdot e_i > 0$ pour tout vecteur e_i de la base canonique. Tous les coefficients diagonaux sont donc strictement positifs, et donc aussi inversibles. On en déduit que la matrice D est inversible et que la méthode de Jacobi est bien définie.
2. (a) Soit λ une valeur propre de J associée au vecteur propre x. On a donc $Jx = \lambda x$, c'est-à-dire $D^{-1}(E+F)x = \lambda x$, soit encore $(E+F)x = \lambda Dx$. On a donc

$$|\sum_{j\neq i} a_{i,j}x_j| = |\lambda||a_{i,i}||x_i|.$$

Soit i tel que $|x_i| = \max_{j=1,N} |x_j|$. Notons que $|x_i| \neq 0$ car x est vecteur propre, donc non nul. En divisant l'égalité précédente par $|x_i|$, on obtient :

$$|\lambda| \leq \sum_{j\neq i} \frac{|a_{i,j}|}{|a_{i,i}|} \leq 1,$$

par hypothèse. On en déduit que $\rho(J) \leq 1$.

(b) Soit $x \in \mathbb{R}^N$ tel que $Jx = \lambda x$ avec $|\lambda| = 1$. Alors

$$|\sum_{j\neq i} a_{i,j}x_j| = |a_{i,i}||x_i|, \text{ pour tout } i = 1, \dots, N. \tag{1.8.84}$$

On a donc

$$\begin{aligned} \sum_{j\neq i} |a_{i,j}||x_i| &\leq |a_{i,i}||x_i| = |\sum_{j\neq i} a_{i,j}x_j| \\ &\leq \sum_{j\neq i} |a_{i,j}||x_j| \text{ pour tout } i = 1, \dots, N. \end{aligned} \tag{1.8.85}$$

Si A est diagonale, alors en vertu de (1.8.84), $x_i = 0$ pour tout $i = 1, \dots, N$. Supposons maintenant A non diagonale. On déduit alors de (1.8.85) que

$$\frac{|x_i|}{|x_j|} \leq 1 \text{ pour tout } i \neq j.$$

Donc $|x_i| = |x_j|$ pour tout i, j.

Comme de plus, par hypothèse,

$$|a_{i_0,i_0}| > \sum_{j \neq i_0} |a_{i_0,j}|,$$

on a donc, si $|x_{i_0}| \neq 0$,

$$\sum_{j \neq i_0} |a_{i_0,j}||x_{i_0}| < |a_{i_0,i_0} x_{i_0}| \leq \sum_{j \neq i_0} |a_{i_0,j} x_{i_0}|,$$

ce qui est impossible. On en déduit que $x = 0$.

On a ainsi prouvé que J n'admet pas de valeur propre de module égal à 1, et donc par la question précédente, $\rho(J) < 1$, ce qui prouve que la méthode converge.

(c) La matrice A de l'exercice 33 est à diagonale fortement dominante. Donc la méthode de Jacobi converge.

Exercice 36 page 48 (Méthode de Jacobi et relaxation)

1. $J = D^{-1}(E + F)$ peut ne pas être symétrique, même si A est symétrique :

   En effet, prenons $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

   Alors

   $$J = D^{-1}(E+F) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix}.$$

   donc J n'est pas symétrique.

2. On applique l'exercice précédent pour l'application linéaire T de matrice D, qui est, par hypothèse, définie positive (et évidemment symétrique puisque diagonale) et $S = E + F$, symétrique car A est symétrique.
   Il existe donc $(f_1 \dots f_N)$ base de E et $(\mu_1 \dots \mu_N) \in \mathbb{R}^N$ tels que

   $$Jf_i = D^{-1}(E+F)f_i = \mu_i f_i, \quad \forall i = 1, \dots, N, \text{ et } (Df_i, f_j) = \delta_{ij}.$$

3. Par définition de J, tous les éléments diagonaux de J sont nuls et donc sa trace également. Or $TrJ = \sum_{i=1}^{N} \mu_i$.

   Si $\mu_i > 0 \ \forall i = 1, \dots, N$, alors $TrJ > 0$, donc $\exists i_0; \ \mu_i \leq 0$ et comme $\mu_1 \leq \mu_{i_0}$, on a $\mu_1 \leq 0$. Un raisonnement similaire montre que $\mu_N \geq 0$.

4. La méthode de Jacobi converge si et seulement si $\rho(J) < 1$ (théorème 1.27 page 28). Or, par la question précédente, $\rho(A) = max(-\mu_1, \mu_N)$. Supposons que $\mu_1 \leq -1$, alors $\mu_1 = -\alpha$, avec $\alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_1$ ou encore $(E+F)f_1 = -\alpha D f_1$, ce qui s'écrit aussi $(D+E+F)f_1 = D(1-\alpha)f_1$ c'est–à–dire $(2D-A)f_1 = \beta D f_1$ avec $\beta \leq 0$. On en déduit que $((2D-A)f_1, f_1) = \beta \leq 0$, ce qui contredit le fait que $2D - A$ est définie positive. En conséquence, on a bien $\mu_1 \geq -1$.

   Supposons maintenant que $\mu_N = \alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_N$, soit encore $(E+F)f_N = -\alpha D f_N$. On en déduit que $Af_N = (D - E - F)f_N = D(1-\alpha)f_N = D\beta f_N$ avec $\beta \leq 0$. On a alors$(Af_N, f_N) \leq 0$, ce qui contredit le fait que A est définie positive.

5. Par définition, on a :$D\tilde{x}^{(n+1)}s = (E+F)x^{(n)} + b$ et $x^{(n+1)} = \omega\tilde{x}^{(n+1)} + (1-\omega)x^{(n)}$. On a donc $x^{(n+1)} = \omega[D^{-1}(E+F)x^{(n)} + D^{-1}b] + (1-\omega)x^{(n)}$ c'est–à–dire $x^{(n+1)} = [Id - \omega(Id - D^{-1}(E+F))]x^{(n)} + \omega D^{-1}b$,, soit encore $\frac{1}{\omega}Dx^{(n+1)} = [\frac{1}{\omega}D - (D - (E+F))]x^{(n)} + b$. On en déduit que $M_\omega x^{(n+1)} = N_\omega x^{(n)} + b$ avec $M_\omega = \frac{1}{\omega}D$ et $N_\omega = \frac{1}{\omega}D - A$.

6. La matrice d'itération est donc maintenant $J_\omega = M_\omega^{-1} N_\omega$ qui est symétrique pour le produit scalaire $(\cdot,\cdot)_{M_\omega}$ donc en reprenant le raisonnement de la question 2, il existe une base $(\tilde{f}_1, \ldots, \tilde{f}_N) \in (\mathbb{R}^N)^N$ et $(\tilde{\mu}_1, \ldots \tilde{\mu}_N) \subset \mathbb{R}^N$ tels que

$$J_\omega \tilde{f}_i = {M_\omega}^{-1} N_\omega \tilde{f}_i = \omega D^{-1} \left(\frac{1}{\omega} D - A\right) \tilde{f}_i = \tilde{\mu}_i \tilde{f}_i, \quad \forall i = 1, \ldots, N,$$

$$\text{et } \frac{1}{\omega} D\tilde{f}_i \cdot \tilde{f}_j = \delta_{ij}, \quad \forall i, = 1, \ldots, N, \, \forall j, = 1, \ldots, N.$$

Supposons $\tilde{\mu}_1 \leq -1$, alors $\tilde{\mu}_1 = -\alpha$, avec $\alpha \geq 1$ et $\omega D^{-1}(\frac{1}{\omega} D - A)\tilde{f}_1 = -\alpha \tilde{f}_1$, ou encore $\frac{1}{\omega} D - A\tilde{f}_1 = -\alpha \frac{1}{\omega} D\tilde{f}_1$. On a donc $\frac{2}{\omega} D - A\tilde{f}_1 = (1-\alpha)\frac{1}{\omega} D\tilde{f}_1$, ce qui entraîne $(\frac{2}{\omega} D - A)\tilde{f}_1 \cdot \tilde{f}_1 \leq 0$. Ceci contredit l'hypothèse $\frac{2}{\omega} D - A$ définie positive.
De même, si $\tilde{\mu}_N \geq 1$, alors $\tilde{\mu}_N = \alpha$ avec $\alpha \geq 1$. On a alors

$$(\frac{1}{\omega} D - A)\tilde{f}_N = \alpha \frac{1}{\omega} D\tilde{f}_N,$$

et donc $A\tilde{f}_N = (1-\alpha)\frac{1}{\omega} D\tilde{f}_N$ ce qui entraîne en particulier que $A\tilde{f}_N \cdot \tilde{f}_N \leq 0$; or ceci contredit l'hypothèse A définie positive.

7. On cherche une condition nécessaire et suffisante pour que

$$\left(\frac{2}{\omega} D - A\right) x \cdot x > 0, \quad \forall x \neq 0, \tag{1.8.86}$$

ce qui est équivalent à

$$\left(\frac{2}{\omega} D - A\right) f_i \cdot f_i > 0, \quad \forall i = 1, \ldots, N, \tag{1.8.87}$$

où les $(f_i)_{i=1,N}$ sont les vecteurs propres de $D^{-1}(E+F)$. En effet, la famille $(f_i)_{i=1,\ldots,N}$ est une base de $\mathbb{R}^N$, et

$$\begin{aligned} \left(\frac{2}{\omega} D - A\right) f_i &= \left(\frac{2}{\omega} D - D + (E+F)\right) f_i \\ &= \left(\frac{2}{\omega} - 1\right) Df_i + \mu_i Df_i \\ &= \left(\frac{2}{\omega} - 1 + \mu_i\right) Df_i. \end{aligned} \tag{1.8.88}$$

On a donc en particulier $\left(\frac{2}{\omega} D - A\right) f_i \cdot f_j = 0$ is $i \neq j$, ce qui prouve que (1.8.86) est équivalent à (1.8.87). De (1.8.87), on déduit, grâce au fait que $(Df_i, f_i) = 1$,

$$\left(\left(\frac{2}{\omega} D - A\right) f_i, f_i\right) = \left(\frac{2}{\omega} - 1 + \mu_i\right).$$

On veut donc que $\frac{2}{\omega} - 1 + \mu_1 > 0$ car $\mu_1 = \inf \mu_i$, c'est–à–dire : $-\frac{2}{\omega} < \mu_1 - 1$, ce qui est équivalent à : $\omega < \dfrac{2}{1-\mu_1}$.

8. La matrice d'itération J_ω s'écrit :

$$J_\omega = \left(\frac{1}{\omega} D\right)^{-1} \left(\frac{1}{\omega} D - A\right) = \omega I_\omega, \text{ avec } I_\omega = D^{-1}(\frac{1}{\omega} D - A).$$

Soit λ une valeur propre de I_ω associée à un vecteur propre u ; alors :

$$D^{-1}\left(\frac{1}{\omega}D - A\right)u = \lambda u, \ i.e. \ \left(\frac{1}{\omega}D - A\right)u = \lambda D u.$$

On en déduit que

$$(D-A)u + \left(\frac{1}{\omega} - 1\right)Du = \lambda D u, \text{ soit encore}$$

$$D^{-1}(E+F)u = \left(1 - \frac{1}{\omega} + \lambda\right)u.$$

Or f_i est vecteur prore de $D^{-1}(E+F)$ associée à la valeur propre μ_i (question 2). On a donc :

$$D^{-1}(E+F)f_i = \mu_i f_i = \left(1 - \frac{1}{\omega} + \lambda\right)f_i,$$

ce qui est vrai si $\mu_i = 1 - \frac{1}{\omega} + \lambda$, c'est–à–dire $\lambda = \mu_i - 1 - \frac{1}{\omega}$. Donc $\mu_i^{(\omega)} = \omega\left(\mu_i - 1 - \frac{1}{\omega}\right)$ est valeur propre de J_ω associée au vecteur propre f_i.
On cherche maintenant à minimiser le rayon spectral

$$\rho(J_\omega) = \sup_i |\omega(\mu_i - 1 - \frac{1}{\omega})|$$

On a

$$\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq \omega(\mu_i - 1 - \frac{1}{\omega}) \leq \omega(\mu_N - 1 - \frac{1}{\omega}),$$

et

$$-\omega(\mu_N - 1 - \frac{1}{\omega}) \leq -\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq -\omega(\mu_i - 1 - \frac{1}{\omega}),$$

donc

$$\rho(J_\omega) = \max\left(|\omega(\mu_N - 1 - \frac{1}{\omega})|, |-\omega(\mu_1 - 1 - \frac{1}{\omega}|)\right)$$

dont le minimum est atteint (voir Figure 1.8) pour

$$\omega(1-\mu_1) - 1 = 1 - \omega(1-\mu_N) \text{ c'est–à–dire } \omega = \frac{2}{2 - \mu_1 - \mu_N}.$$

Exercice 37 page 49 (Jacobi et Gauss-Seidel pour une matrice tridiagonale)

Corrigé en cours de rédaction

Exercice 39 page 50 (Méthode de Jacobi pour des matrices particulières)

1. Soit $x \in \mathbb{R}^N$, supposons que

$$\|x\|_A = \sum_{i=1}^{N} a_{i,i}|x_i| = 0.$$

Comme $a_{i,i} > 0, \forall i = 1, \ldots, N$, on en déduit que $x_i = 0, \forall i = 1, \ldots, N$. D'autre part, il est immédiat de voir que $\|x+y\|_A \leq \|x\| + \|y\|_A$ pour tout $(x,y) \in \mathbb{R}^N \times \mathbb{R}^N$ et que $\|\lambda x\|_A = |\lambda|\|x\|_A$ pour tout $(x,\lambda) \in \mathbb{R}^N \times \mathbb{R}$. On en déduit que $\|\cdot\|_A$ est une norme sur $\mathbb{R}^N$.

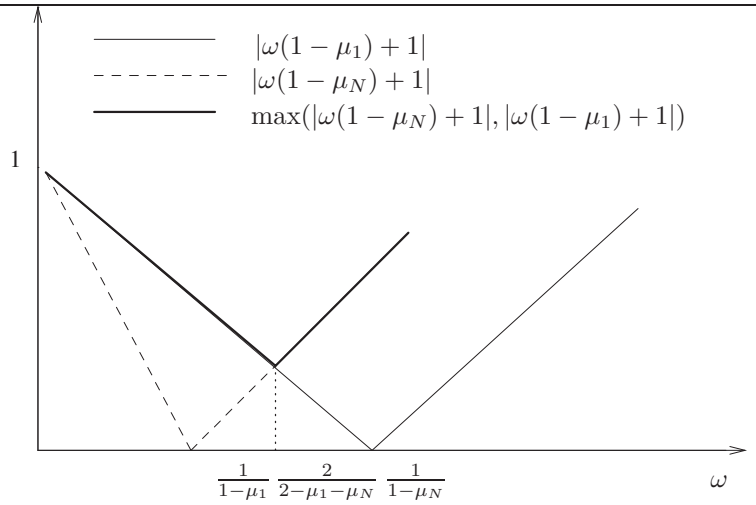

FIGURE 1.6 – Détermination de la valeur de ω réalisant le minimum du rayon spectral.

2. Posons $\tilde{A} = \lambda Id + A$ et notons $\tilde{a}_{i,j}$ ses coefficients. Comme $\lambda \in \mathbb{R}_+^*$,et grâce aux hypothèses (1.6.57)–(1.6.59), ceux-ci vérifient :

$$\tilde{a}_{i,j} \leq 0, \forall i, j = 1, \ldots, N, \, i \neq j, \tag{1.8.89}$$

$$\tilde{a}_{i,i} > \sum_{\substack{i=1 \\ j \neq i}}^{N} a_{i,j}, \forall i = 1, \ldots, N. \tag{1.8.90}$$

La matrice $\tilde{A}$ est donc à diagonale dominante stricte, et par l'exercice 32 page 47, elle est donc inversible.

3. La méthode de Jacobi pour la résolution du système (1.6.60) s'écrit :

$$\tilde{D}u^{(k+1)} = (E+F)u^{(k)} + b, \tag{1.8.91}$$

avec $\tilde{D} = \lambda Id + D$, et $A = D - E - F$ est la décomposition habituelle de A en partie diagonale, triangulaire inférieure et triangulaire supérieure. Comme $a_{i,i} \geq 0$ et $\lambda \in \mathbb{R}_+^*$, on en déduit que $\tilde{D}$ est inversible, et que donc la suite $(u^{(k)})_{k \in \mathbb{N}}$ est bien définie dans $\mathbb{R}$.

4. Par définition de la méthode de Jacobi, on a :

$$u_i^{(k+1)} = \frac{1}{a_{i,i} + \lambda} \Big(\sum_{\substack{j=1,N \\ j \neq i}} a_{i,j} u_j^{(k)} + b_i \Big).$$

On en déduit que

$$u_i^{(k+1)} - u_i^{(k)} = \frac{1}{a_{i,i} + \lambda} \sum_{\substack{j=1,N \\ j \neq i}} a_{i,j} (u_j^{(k)} - u_j^{(k-1)}).$$

et donc

$$\|u^{(k+1)} - u^{(k)}\|_A \leq \sum_{i=1}^{N} \frac{a_{i,i}}{a_{i,i} + \lambda} \sum_{\substack{j=1,N \\ j \neq i}} a_{i,j} (u_j^{(k)} - u_j^{(k-1)}).$$

Or $\frac{a_{i,i}}{a_{i,i}+\lambda} \leq \frac{1}{1+\frac{\lambda}{a_{i,i}}} \leq \frac{1}{1+\alpha}$. On a donc

$$\|u^{(k+1)} - u^{(k)}\|_A \leq \frac{1}{1+\alpha} \sum_{j=1}^{N} (u_j^{(k)} - u_j^{(k-1)}) \sum_{\substack{j=1,N \\ j\neq i}} a_{i,j}.$$

Et par hypothèse, $\sum_{\substack{j=1,N \\ j\neq i}} a_{i,j} = a_{j,j}$. On en déduit que

$$\|u^{(k+1)} - u^{(k)}\|_A \leq \frac{1}{1+\alpha} \|u^{(k)} - u^{(k-1)}\|_A.$$

On en déduit le résultat par une récurrence immédiate.

5. Soient p et $q = p + m \in \mathbb{N}$, avec $m \geq 0$. Par le résultat de la question précédente, on a :

$$\begin{aligned} \|u^{(q)} - u^{(p)}\|_A &\leq \sum_{i=1}^{m} \|u^{(p+i)} - u^{(p-i-1)}\|_A \\ &\leq \|u^{(1)} - u^{(0)}\|_A (\frac{1}{1+\alpha})^p \sum_{i=0}^{m} (\frac{1}{1+\alpha})^i \end{aligned}$$

Or $\alpha > 0$ donc la série de terme général $(\frac{1}{1+\alpha})^i$, et on a :

$$\begin{aligned} \|u^{(q)} - u^{(p)}\|_A &\leq \|u^{(1)} - u^{(0)}\|_A (\frac{1}{1+\alpha})^p \sum_{i=0}^{+\infty} (\frac{1}{1+\alpha})^i \\ &\leq (1+\frac{1}{\alpha}) \|u^{(1)} - u^{(0)}\|_A (\frac{1}{1+\alpha})^p \\ &\to 0 \text{ lorsque } p \to +\infty. \end{aligned}$$

On en déduit que pour tout $\epsilon > 0$, il existe N tel que si $p, q > N$ alors $\|u^{(q)} - u^{(p)}\|_A \leq \epsilon$, ce qui montre que la suite est de Cauchy, et donc qu'elle converge. Soit $\overline{u}$ sa limite. En passant à la limite dans (1.8.91), on obtient que $\overline{u}$ est solution de (1.6.60).

Exercice 41 page 51 (Une méthode itérative particulière)

1. Det $(A) = -1$ et donc A est inversible.

2. Det $(\frac{1}{\omega} Id - E) = \frac{1}{\omega} \left(\frac{1}{\omega^2} - 2 \right)$. Or $\omega \in]0, 2[$. Donc la matrice $\frac{1}{\omega} Id - E$ est inversible si $\omega \neq \frac{\sqrt{2}}{2}$.

3. Les valeurs propres de $\mathcal{L}_\omega$ sont les complexes λ tels qu'il existe $x \in C^3, x \neq 0$, t.q : $\mathcal{L}_\omega x = \lambda x$, c'est–à–dire :

$$\left(F + \frac{1-\omega}{\omega} Id \right) x = \lambda \left(\frac{1}{\omega} Id - E \right) x,$$

soit encore $M_{\lambda,\omega} x = 0$, avec $M_{\lambda,\omega} = \omega F + \lambda\omega E + (1 - \omega - \lambda) Id$.

Or

$$\begin{aligned} Det\,(M_{\lambda,\omega}) &= (1-\omega-\lambda)((1-\omega-\lambda)^2 - 2\lambda^2\omega^2) \\ &= (1-\omega-\lambda)(1-\omega-(1+\sqrt{2}\omega)\lambda)(1-\omega-(1-\sqrt{2}\omega)\lambda) \end{aligned}$$

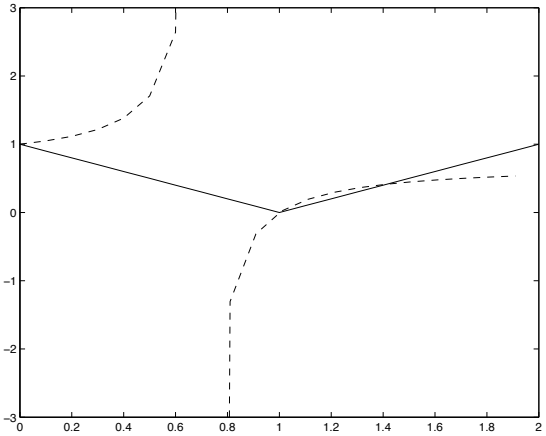

FIGURE 1.7 – Graphe des valeurs propres λ_1 et λ_3

Les valeurs propres de $\mathcal{L}_\omega$ sont donc réelles, et égales à

$$\lambda_1 = 1-\omega, \lambda_2 = \frac{1-\omega}{1+\sqrt{2}\omega} \text{ et } \lambda_3 = \frac{1-\omega}{1-\sqrt{2}\omega}.$$

Par définition, le rayon spectral $\rho(\mathcal{L}_\omega)$ de la matrice $\mathcal{L}_\omega$ est égal à $\max(|\lambda_1|, |\lambda_2|, |\lambda_3|)$. Remarquons tout d'abord que $|1+\sqrt{2}\omega| > 1, \forall \omega \in]0,2[$, et donc $|\lambda_1| > |\lambda_2|, \forall \omega \in]0,2[$. Il ne reste donc plus qu'à comparer $|\lambda_1|$ et $|\lambda_3|$. Une rapide étude des fonctions $|\lambda_1|$ et $|\lambda_3|$ permet d'établir le graphe représentatif ci-contre.
On a donc :

$$\rho(\mathcal{L}_\omega) = |\lambda_3(\omega)| = \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right| \text{ si } \omega \in]0, \sqrt{2}]$$

$$\rho(\mathcal{L}_\omega) = \lambda_1(\omega) = |1-\omega| \text{ si } \omega \in [\sqrt{2}, 2[.$$

4. La méthode est convergente si $\rho(\mathcal{L}_\omega) < 1$; Si $\omega \in [\sqrt{2}, 2[, \rho(\mathcal{L}_\omega) = \omega - 1 < 1$; si $\omega \in]0, \sqrt{2}[$,

$$\rho(\mathcal{L}_\omega) = \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right| < 1$$

dès que $\dfrac{1-\omega}{\sqrt{2}\omega - 1} < 1$, c'est à dire $\omega > \dfrac{2}{1+\sqrt{2}}$.
Le minimum de $\rho(\mathcal{L}_\omega)$ est atteint pour $\omega_0 = 1$, on a alors $\rho(\mathcal{L}_\omega) = 0$.

Exercice 43 page 52 (Méthode de la puissance pour calculer le rayon spectral de A)

1. Comme A est une matrice symétrique, A est diagonalisable dans $\mathbb{R}$. Soit $(f_1, \ldots, f_N) \in (\mathbb{R}^N)^N$ une base orthonormée de vecteurs propres de A associée aux valeurs propres $(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N$. On décompose $x^{(0)}$ sur $(f_i)_{i=1,\ldots,N}$: $x^{(0)} = \sum_{i=1}^N \alpha_i f_i$. On a donc $Ax^{(0)} = \sum_{i=1}^N \lambda_i \alpha_i f_i$ et $A^n x^{(0)} = \sum_{i=1}^N \lambda_i^n \alpha_i f_i$.
   On en déduit :

$$\frac{x^{(n)}}{\lambda_N^n} = \sum_{i=1}^N \left(\frac{\lambda_i}{\lambda_N}\right)^n \alpha_i f_i.$$

Comme $-\lambda_N$ n'est pas valeur propre,

$$\lim_{n\to+\infty}(\frac{\lambda_i}{\lambda_N})^n = 0 \text{ si } \lambda_i \neq \lambda_N. \tag{1.8.92}$$

Soient $\lambda_1, \ldots, \lambda_p$ les valeurs propres différentes de λ_N, et $\lambda_{p+1}, \ldots, \lambda_N = \lambda_N$. On a donc
$\lim_{n\to+\infty} \frac{x^{(n)}}{\lambda_N^n} = \sum_{i=p+1}^N \alpha_i f_i = x$, avec $Ax = \lambda_N x$.
De plus, $x \neq 0$: en effet, $x^{(0)} \notin (Ker(A - \lambda_N Id))^\perp = Vect\{f_1, \ldots, f_p\}$, et donc il existe $i \in \{p+1, \ldots, N\}$ tel que $\alpha_i \neq 0$.
Pour montrer (b), remarquons que :

$$\|x^{(n+1)}\| = \sum_{i=1}^N \lambda_i^{n+1}\alpha_i \text{ et } \|x^{(n)}\| = \sum_{i=1}^N \lambda_i^n \alpha_i$$

car $(f_1, \ldots, f_N)$ est une base orthonormée. On a donc

$$\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} = \lambda_N^n \frac{\|\frac{x^{(n+1)}}{\lambda_N^{n+1}}\|}{\|\frac{x^{(n)}}{\lambda_N}\|} \to \lambda_N \frac{\|x\|}{\|x\|} = \lambda_N \text{ lorsque } n \to +\infty.$$

2. a) La méthode I s'écrit à partir de $x^{(0)}$ connu : $x^{(n+1)} = Bx^{(n)} + c$ pour $n \geq 1$, avec $c = (I - B)A^{-1}b$.
On a donc

$$\begin{aligned} x^{(n+1)} - x &= Bx^{(n)} + (Id - B)x - x \\ &= B(x^{(n)} - x). \end{aligned} \tag{1.8.93}$$

Si $y^{(n)} = x^{(n)} - x$, on a donc $y^{(n+1)} = By^{(n)}$, et d'après la question 1a) si $y^{(0)} \not\perp \mathrm{Ker}(B - \mu_N Id)$ où μ_N est la plus grande valeur propre de B, (avec $|\mu_N| = \rho(B) et - \mu_N$ non valeur propre), alors

$$\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty,$$

c'est–à–dire

$$\frac{\|x^{(n+1)} - x\|}{\|x^{(n)} - x\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty.$$

b) On applique maintenant 1a) à $y^{(n)} = x^{(n+1)} - x^{(n)}$ avec

$$y^{(0)} = x^{(1)} - x^{(0)} \text{ où } x^{(1)} = Ax^{(0)}.$$

On demande que $x^{(1)} - x^{(0)} \notin \mathrm{Ker}(B - \mu_N Id)^\perp$ comme en a), et on a bien $y^{(n+1)} = By^{(n)}$, donc $\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B)$ lorsque $n \to +\infty$.

Exercice 45 page 53 (Méthode QR pour la recherche de valeurs propres)

En cours de rédaction.